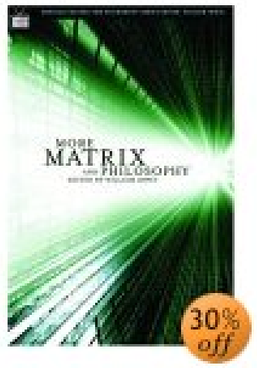
Why Make a Matrix? And Why You Might Be In One



Nick Bostrom

Faculty of Philosophy, Oxford University

www.nickbostrom.com

Forthcoming in *More Matrix and Philosophy: Revolutions and Reloaded Decoded*, ed. William Irwin (Open Court, 2005)

The Purpose of the Matrix

Why the Matrix? Why did the machines do it? (Human brains may be many things, but efficient batteries they are not.) How could they justify a world whose inhabitants are systematically deceived about their fundamental reality, ignorant about the reason why they exist, and subject to all the cruelty and suffering that we witness in the world around us? Children dying of AIDS; lovers separated by war and poverty; cancer patients tormented by unbearable pain; stroke victims deprived of their use of language and reason… One would think nobody but a sadist could have the imagination to think up these horrors, much less possess the desire to create a world that contains them in such abundance. But the machines did it, at least that's how the story goes.

Although the world of the Matrix they created is far from perfect, it is – arguably – better than no world at all, the elimination of all human beings. Still, the machines could have created a world containing much more goodness, happiness, wisdom, personal growth, love and beauty, a world that was free of most of the natural and manmade evil that pervades our world. Indeed, as the story goes, they tried that, but supposedly it didn't work.

> *Agent Smith*: Did you know that the first Matrix was designed to be a perfect human world. Where none suffered. Where everyone would be happy. It was a disaster. No one would accept the program. Entire crops were lost. Some believed that we lacked the programming language to describe your perfect world. But I believe that as a species, human beings define their reality through misery and suffering. The perfect world *was a dream* that your primitive cerebrum kept trying to wake up from. Which is why the Matrix was redesigned to this, the peak of your civilization

The existence of unnecessary evil is one of the most powerful arguments against the belief that the world was created by an all-powerful, all-knowing, and perfectly good God. Theologians have spent centuries trying to answer it, and with very questionable success. But the problem of evil is only a problem if one assumes that the world was created by an omnipotent and perfectly good being. If one assumes instead that the creator was not perfectly good, and perhaps not even omnipotent, then it would be much easier to reconcile the view that our world was created with its seemingly obvious ethical shortcomings.

What about you? You're not all-powerful, all-knowing, and perfectly good. But what if you had the ability to create this kind of Matrix, would you do it? Even if *you* would not have chosen to create a world like this, there are many other people who do not share your scruples. If these people had the ability to create Matrices, some of their works might well look like the world in which we find ourselves.

Why might they choose to build a Matrix like our reality? One can think of many possible reasons —setting aside the daft idea of using human brains as batteries. But perhaps future historians would create a Matrix that mimicked the history of their own species. They might do this to find out more about their past, or to explore counterfactual historical scenarios. In the world of the Architect(s), Napoleon may have succeeded in conquering Europe, and our world might be a Matrix created to research what would have happened if Napoleon had been defeated. Or perhaps there will be future artists who create Matrices as an art form much like we create movies and operas. Or perhaps the tourist industry will create simulations of interesting historical epochs so that their contemporaries can go on themed holidays to some bygone age by entering into the simulation and interacting with its inhabitants. The possible motives are myriad, and if future people are anything like present people, and if they have the technological might and the legal right to create Matrices, we would expect that many Matrices would be created, including ones that would look like the world that we are experiencing.

The Simulation Argument

If each advanced civilization created many Matrices of their own history, then most people like us, who live in a technologically more primitive age, would live inside Matrices rather than outside them. If this were the case, where would you most likely be?

The so-called Simulation argument, which I introduced a few years ago, makes this line of reasoning more precise and takes it to its logical conclusion. The conclusion is that there are three basic possibilities at least one of which is true. The first possibility is that the human species will almost certainly go extinct before becoming technologically mature. The second possibility is that almost no technologically mature civilization is interested in building Matrices. The third possibility is that we are almost certainly living in a Matrix. Why? Because if the first two possibilities are not

the case, then there are more “people” living in Matrices than in “real worlds.” As a “person” then the chances are that you are living in a Matrix rather than in a “real world.”

The Simulation argument does not tell us which of these three possibilities obtain, only that at least one of them does. The argument employs some math and probability theory, but the basic idea can be understood without recourse to technical apparatus.[1]

Building a Matrix

Creating comprehensive Matrices that are indistinguishable from non-simulated reality is, of course, far beyond our current technological capability. Even so, we can estimate the computational requirements for creating such virtual realities.

Rather than confining the construction project to creating a virtual reality simulation, we can consider a more ambitious project that also involves the creation of the inhabitants of the Matrix. Instead of having pink gooey pods with biological humans floating in them being fed sensory input from a simulated reality, it would be more efficient to replace the brains with simulations of brains. Many philosophers and cognitive scientists believe that such brain-simulations would be conscious, provided the simulation was sufficiently detailed and accurate.

Estimates of the human brain’s computational power have been given and estimates of the computational power that would be available to a technologically mature civilization can also been made. While these estimates are very approximate, it turns out that even when allowing for a large margin of error, the computational resources of a mature civilization would suffice to create very many Matrices. Even a single planetary-sized computer, constructed with advanced molecular nanotechnology, could simulate the entire mental history of humankind by using less than one millionth of its computing power for one second; and this presupposes only already known computational mechanisms and engineering principles. A single civilization may eventually build millions of such computers. We can conclude that a technologically mature civilization would have enough computing power such that even if it devoted but a tiny fraction of it to creating Matrices, there would soon be many more simulated people than there were people living in the original history of that civilization.

These simulations would not have to be perfect. They would only have to be good enough to fool its inhabitants. It would not be necessary to simulate every object down to the subatomic level (something that would definitely be infeasible). If the book you are holding in your hands is a simulated book, the simulation would only need to include its visual appearance, its weight and texture, and a few other macroscopic properties, because you have no way of knowing what its individual atoms are doing at this moment. If you were to study the book more carefully, for example by examining it under a powerful microscope, additional details of the simulation could be

filled in as needed. Objects that nobody is perceiving could have an even more compressed representation. Such simplifications would dramatically reduce the computational requirements.

Three Possibilities

Given that the Architects of a technologically mature civilization could create a vast number of Matrices even by devoting just a small fraction of their resources to that end, an interesting implication follows. Consider the set of civilizations that are at similar level of technological development as our own current civilization. Suppose that some non-trivial fraction of these eventually go on to become technologically mature. Suppose, furthermore, that some non-trivial fraction of these devote a non-negligible proportion of their resources to building Matrices. Then most people like us live in Matrices rather than outside them. There are thus three basic possibilities: *either* almost every civilization like ours go extinct before reaching technological maturity; *or* almost every mature civilization lacks any interest in building Matrices; *or* almost all people with our kind of experiences live in Matrices.

Let us think a little about these three possibilities. If almost every civilization at our current stage goes extinct before becoming technologically mature, then our future looks relatively bleak. For if such a premature ending were the fate awaiting most civilizations, we would have to suspect that the same will hold for our civilization in particular. This is because we seem to lack any reason for thinking that our civilization will be luckier than most other civilizations at our stage.

The second possibility is less depressing. It might turn out that almost all technologically mature civilizations lose interest in building Matrices. Maybe the potential Architects of the future will not share any of the possible motives for building Matrices that we discussed above. Presumably, Architects would have used their advanced technology to improve their own capacities, so they may be superintelligent and have complete control over their own mental states. Rather than resorting to Matrix-building for recreation, they may obtain pleasure more efficiently by direct stimulation of their brains' pleasure centers. Their science may be so advanced that they have little to learn from running simulations of their historical past. Furthermore, they might develop ethical norms that prohibit the creation of Matrices. So we cannot infer from the fact that many *current* people would be tempted to construct Matrices that the same would hold for the super-advanced folks that would actually have the ability to act on this motive.

The third possibility is the most intriguing. If the vast majority of all people with other kind of experiences live in Matrices then *we* probably live in a Matrix. Unless we had some specific evidence to the contrary, we would therefore have to conclude that the world we see around us exists only by virtue of being simulated on a powerful computer built by some technologically highly advanced Architect.

Not the Old Brain-in-a-Vat Argument

For hundreds of years, philosophers have pondered the question how we can know that the external world exists. Descartes (1596-1650) posed this question in his *Meditationes*, and considered the scenario where a hypothetical evil demon caused us to have erroneous beliefs about external objects. In more recent years, Descartes' skeptical scenario has been given a more modern finish, and instead of a demon one is now asked to imagine a mad scientist who has extracted one's brain and who keeps it in a vat where the scientist is stimulating it with electrical signals replicating the sensory input that the brain would have had if it had interacted with a very different environment from that which is present in the real world. This is, of course, is the predicament explored in the Matrix movie. How can one possibly know that one is not such a brain in a vat, the philosophical skeptic challenges, given that all the appearances we experience could be the experiences of an envatted brain?

The argument outlined above provides a much stronger reason for taking seriously the possibility that we are living in a Matrix. The traditional skeptical argument offers no positive ground for thinking that we are living in a Matrix. At best, it shows that we cannot completely rule out that possibility, but we remain free to assign it a very small or negligible probability. If there are no mad scientists who experiment on conscious envatted human brains, then we are not envatted. Even if there were a few such brains-in-vats, they might be extremely rare compared to the brains-in-crania that interact with the external world in the normal way; and if so, then it may be highly unlikely that we would be among the envatted ones.

The Simulation argument, by contrast, adopts as its starting point that things are the way they seem to be and that science gives us reliable information about the world. Part of this information concerns the technological capabilities that an advanced civilization would be able to develop. Among these would be the capability to create Matrices. Crucially, it seems that they could easily create Matrices in astronomical numbers. From this we can then conclude that *either* technologically mature civilizations that are interested in creating Matrices are extremely rare compared to civilizations at our own current stage of development *or* almost all people like us live in Matrices. And from this, the division into three the three basic possibilities mentioned above follows.

The Simulation argument itself doesn't tell us which one of these three possibilities obtain. In fact, we do not currently have any strong evidence either for or against either of these three possibilities. We should therefore assign them all a significant probability. In particular, we should take seriously the possibility that we are living in a Matrix. We might still think that the probability is less than 50%. A degree of belief of something like 20% would seem quite reasonable given our current information.

How Could You Tell If You Are In A Matrix?

Consider the predicament of Neo and his fellow rebels in the trilogy. They *know* there are many Matrices. They lead parts of their lives inside a Matrix. They know that most of their compatriots spend their whole lives in a Matrix. Given this, they should be extremely reluctant to think that they have escaped their Matrix. What appears to be an escape could easily just be simulated escape, so that they exit one level of the Matrix only to reemerge at another. The Wachowski brothers can of course stipulate that this is not the case and that the heroes really do get to experience "real" reality. But if Neo were rational, he would never be able to be at all confident that this is what happens.

If the Wachowski brothers had created a *real* Matrix (rather than just a movie *about* a Matrix), then, if they were rational, they would have to conclude that *they* themselves are almost certainly in a Matrix. For if we develop the capability to create our own Matrices, and if we decide to make use of this capability, we would obtain very strong evidence against the first two possibilities: that it is *not* the case that almost all civilizations at our current stage go extinct before reaching technologically maturity and that it is *not* the case that almost all mature civilizations lose interest in creating Matrices. This would leave us with only the third possibility—that we almost certainly inhabit a Matrix.

But what about the situation we actually find ourselves in? The Simulation argument aside, would it be possible to detect any direct signs of being in a Matrix? Is there a kind of "splinter in the mind" that would indicate that all is not right with reality? Certainly, if the Architects of a Matrix wished to reveal themselves, it would be easy enough for them to do so. For example, they could make a window pop up in our visual field with the text "YOU ARE LIVING IN A MATRIX. CLICK HERE FOR MORE INFORMATION".

In the movie, the Oracle tells us that UFOs, Ghosts, and other strange sights are the manifestations of malfunctions in the Matrix that are being covered up.

> *The Oracle*: Look, see those birds? At some point a program was written to govern them. A program was written to watch over the trees, and the wind, the sunrise, and sunset. There are programs running all over the place. The ones doing their job, doing what they were meant to do, are invisible. You'd never even know they were here. But the other ones, well, we hear about them all the time.
>
> *Neo*: I've never heard of them.
>
> *The Oracle*: Of course you have. Every time you've heard someone say they saw a ghost, or an angel. Every story you've ever heard about vampires, werewolves, or aliens is the system assimilating some program that's doing something they're not supposed to be doing.

Déjà vu is a sign of a glitch in the Matrix, which is re-running a sequence to cover something that has changed. Some people have written to me that they have found signs that we are in a Matrix. One person, for instance, told me that he could see flickering pixels when he looked in his bathroom mirror. Another person wrote that he could hear voices in his head. But even if we are in a Matrix, it is far more likely that such phenomena are the result of imperfections in the reporters rather than in the Matrix itself. There are many perfectly ordinary explanations for why some people should report having these kinds of experiences, including mental illness, over-excited imagination, gullibility, and so forth. Dysfunctional brains could be simulated just as easily as properly functioning ones, and including them in the simulation may indeed add to its verisimilitude.

Building any kind of Matrix at all that contains conscious simulated brains would be tremendously difficult. Any being capable of such a feat would almost certainly also be able to prevent any glitches in their Matrix from being noticed by its inhabitants. Even if some people did notice an anomaly, the Architect could backtrack the simulation a few seconds and rerun it in a way that avoided the anomaly entirely or else could simply edit out the memory of the anomaly from whoever had noticed something suspect.

How To Live In A Matrix

If we knew the Architects' motives for designing Matrices then the hypothesis that we live in one might have major practical consequences. But in fact we know almost nothing about what these motives might be. Because of this ignorance, our best method for getting around in our Matrix (if that is where we are) is to study the patterns we find in the world we experience. We would run experiments, discover regularities, build models, and extrapolate from past events. In other words, we would apply the scientific method and common sense in the same way as if we knew that we were not in a Matrix. To a first approximation, therefore, the answer to how you should live if you are in a Matrix is that you should live the same way as if you are not in a Matrix.

The Simulation argument does, however, have some more subtle practical ramifications, even if we set aside the other two possibilities to which it points (which do not entail that we are in a Matrix). Some scenarios that would otherwise seem to have been foreclosed by our current scientific understanding again become real possibilities if we inhabit a Matrix. For instance, while the physical world cannot suddenly pop out of existence, a simulated reality could do so at any time if the Architect decides to pull the plug. An afterlife would also be a real possibility. When a person dies in a simulation, he or she could be resurrected in another simulation, or the Architect could uplift the deceased into his own level of reality.

It is also conceivable that only some people are simulated in enough detail to be conscious while others may be simulated at a cruder level allowing them to appear and behave much like the real people but without having any subjective experience. The so-called "problem of other minds"—

how we can know that other people are really conscious and are not just behaving as if they were—is another old chestnut of philosophy. There is, however, no consensus that such "zombie" people are possible even in principle. Some people have argued that it is necessarily true that anybody who acts sufficiently like a normal human being must also have conscious experience. (Whether this view would entail that your least favorite politicians cannot be zombies is a question on which more research is required.)

Another possibility it that the Architect might decide to reward or punish his simulated creatures, perhaps on the basis of moral criteria. If you might be in a Matrix, this consideration may give you a novel self-interested reason for behaving morally. The situation would be analogous to the case where God is watching and judging you except that the role of the final judge would not be a supernatural being but the physical person or persons who built the Matrix.

It would be misleading to say that if we are in a Matrix then we and the world around us do not really exist. It would be more accurate to say that the reality of these things is of a somewhat different nature than we thought before. Your nose would still be real; only, its reality would consist in being simulated on a powerful computer. The computer and the electrical activity of its circuitry would be physical phenomena in the more basic level of reality inhabited by the Architect of the Matrix.

Matrices Repeated and Stacked

When Neo stopped the Sentinels with his mind outside the Matrix at the end of *Reloaded* the speculation began. Was there a Matrix on top of the Matrix? As *Revolutions* revealed, there was not. But there could have been. A mature civilization would have enough computing power to run astronomically many Matrices. If we are in a Matrix, therefore, there are probably vast numbers of other Matrices, which differ from ours in some detail or in their overall design. These other Matrices may be run sequentially, as in the movie, or simultaneously by time-sharing the same processor or by using multiple computers. From the viewpoint of the simulated inhabitants, it makes little difference how the Matrices are implemented.

A Matrix may contain a civilization that matures and proceeds to build its own Matrices in the simulation. Realty could thus contain many levels, with computers being simulated inside computers which are themselves simulated, and so forth. How many layers of simulation there could be depends on the computing power available to the bottom-level Architect (who is not simulated). Since all the higher levels of simulation would ultimately be implemented on this Architect's computer, he would have to shoulder the cost of all the simulations and all the simulated people. If his computing power is limited, there may be only a small number of levels.

As we noted above, all Architects would have strong reason to think that they themselves might be in a Matrix. (If the Architects at the basement level believed this they would be mistaken, but only because of bad epistemic luck, not because of any fault in their reasoning.) If we combine this insight with the speculation that moral considerations may play a part in determining the treatment some simulated people receive at the hands of their Architects, we are led to the peculiar thought that everybody—not just the simulated people—may have a self-interested reason for behaving morally. If behaving morally towards somebody includes judging and treating them according to moral criteria, this could further strengthen the reason that everybody have for behaving morally. The stronger that reason is, the more we would expect that people would be motivated by it. And the more people are likely to be motivated to treating their simulated creatures morally, the stronger this reason would become. This reasoning can be iterated indefinitely in a truly "virtuous circle," albeit a rather tenuous one as it relies not only on the possibility that we are in a simulation but also on tenuous speculations about the motives of the Architects.

At a minimum, the Simulation argument provides many exciting avenues for philosophical thinking. But if it is sound—and so far it has not been refuted—it could also provide various suggestions, however tentative and ambiguous, for how we should go about our lives and for what we should expect in the future. When we follow through the logical implications of what we think we know, we discover just how much we don't yet know.

---

[1] For the full story, see (2003) "Are You Living In A Computer Simulation?" Philosophical Quarterly. Vol. 53, No. 211, pp. 243-255. This and other related papers are available at simulation-argument.com.