1 Brains in Vats

*The Matrix* presents a version of an old philosophical fable: the brain in a vat. A disembodied brain is floating in a vat, inside a scientist’s laboratory. The scientist has arranged that the brain will be stimulated with the same sort of inputs that a normal embodied brain receives. To do this, the brain is connected to a giant computer simulation of a world. The simulation determines which inputs the brain receives. When the brain produces outputs, these are fed back into the simulation. The internal state of the brain is just like that of a normal brain, despite the fact that it lacks a body. From the brain’s point of view, things seem very much as they seem to you and me.

The brain is massively deluded, it seems. It has all sorts of false beliefs about the world. It believes that it has a body, but it has no body. It believes that it is walking outside in the sunlight, but in fact it is inside a dark lab. It believes it is one place, when in fact it may be somewhere quite different. Perhaps it thinks it is in Tucson, when it is actually in Australia, or even in outer space.

Neo’s situation at the beginning of *The Matrix* is something like this. He thinks that he lives in a city, he thinks that he has hair, he thinks it is 1999, and he thinks that it is sunny outside. In reality, he is floating in space, he has no hair, the year is around 2199, and the world has been darkened by war. There are a few small differences from the vat scenario above: Neo’s brain is located in a body, and the computer simulation is controlled by machines rather than by a scientist. But the essential details are much the same. In effect, Neo is a brain in a vat.

Let’s say that a matrix (lower-case “m”) is an artificially-designed computer simulation of a world. So the Matrix in the movie is one example of a matrix. And let’s say that someone is envatted, or that they are in a matrix, if they have a cognitive system which receives its inputs from and sends its outputs to a matrix. Then the brain at the beginning is envatted, and so is Neo.

We can imagine that a matrix simulates the entire physics of a world, keeping track of every last particle throughout space and time. (Later, we will look at ways in which this set-up might be varied.) An envatted being will be associated with a particular simulated body. A connection is arranged so that whenever this body receives sensory inputs inside the simulation, the
envatted cognitive system will receive sensory inputs of the same sort. When the envatted
cognitive system produces motor outputs, corresponding outputs will be fed to the motor
organs of the simulated body.

When the possibility of a matrix is raised, a question immediately follows. How do I know that
I am not in a matrix? After all, there could be a brain in a vat structured exactly like my brain,
hooked up to a matrix, with experiences indistinguishable from those I am having now. From
the inside, there is no way to tell for sure that I am not in the situation of the brain in a vat. So
it seems that there is no way to know for sure that I am not in a matrix.

Let us call the hypothesis that I am in a matrix and have always been in a matrix the *Matrix
Hypothesis*. Equivalently, the Matrix Hypothesis says that I am envatted and have always been
envatted. This is not quite equivalent to the hypothesis that I am in the Matrix, as the Matrix is
just one specific version of a matrix. For now, I will ignore the some complications that are
specific to the Matrix in the movie, such as the fact that people sometimes travel back and forth
between the Matrix and the external world. These issues aside, we can think of the Matrix
Hypothesis informally as saying that I am in the same sort of situation as people who have
always been in the Matrix.

The Matrix Hypothesis is one that we should take seriously. As Nick Bostrom has suggested,
it is not out of the question that in the history of the universe, technology will evolve that will
allow beings to create computer simulations of entire worlds. There may well be vast numbers
of such computer simulations, compared to just one real world. If so, there may well be many
more beings who are in a matrix than beings who are not. Given all this, one might even infer
that it is more likely that we are in a matrix than that we are not. Whether this is right or not,
it certainly seems that we cannot be certain that we are not in a matrix.

Serious consequences seem to follow. My envatted counterpart seems to be massively deluded.
It thinks it is in Tucson; it thinks it is sitting at a desk writing an article; it thinks it has a body.
But on the face of it, all of these beliefs are false. Likewise, it seems that if I am envatted, my
own corresponding beliefs are false. If I am envatted, I am not really in Tucson, I am not really
sitting at a desk, and I may not even have a body. So if I don’t know that I am not envatted,
then I don’t know that I am in Tucson, I don’t know that I am sitting at a desk, and I don’t
know that I have a body.

The Matrix Hypothesis threatens to undercut almost everything I know. It seems to be a
skeptical hypothesis: a hypothesis that I cannot rule out, and one that would falsify most of my
beliefs if it were true. Where there is a skeptical hypothesis, it looks like none of these beliefs
count as genuine knowledge. Of course the beliefs might be true — I might be lucky, and not
be envatted — but I can’t rule out the possibility that they are false. So a skeptical hypothesis
leads to skepticism about these beliefs: I believe these things, but I do not know them.

To sum up the reasoning: I don’t know that I’m not in a matrix. If I’m in a matrix, I’m probably
not in Tucson. So if I don’t know that I’m not in a matrix, then I don’t know that I’m in Tucson.
The same goes for almost everything else I think I know about the external world.
2 Envatment Reconsidered

This is a standard way of thinking about the vat scenario. It seems that this view is also endorsed by the people who created *The Matrix*. On the DVD case for the movie, one sees the following:

**Perception:** Our day-in, day-out world is real.

**Reality:** That world is a hoax, an elaborate deception spun by all-powerful machines that control us. Whoa.

I think this view is not quite right. I think that even if I am in a matrix, my world is perfectly real. A brain in a vat is not massively deluded (at least if it has always been in the vat). Neo does not have massively false beliefs about the external world. Instead, envatted beings have largely *correct* beliefs about their world. If so, the Matrix Hypothesis is not a skeptical hypothesis, and its possibility does not undercut everything that I think I know.

Philosophers have held this sort of view before. The 18th-century philosopher George Berkeley held, in effect, that appearance is reality. (Recall Morpheus: “What is real? How do you define real? If you’re talking about what you can feel, what you can smell, what you can taste and see, then real is simply electrical signals interpreted by your brain.”) If this is right, then the world perceived by envatted beings is perfectly real: they have all the right appearances, and appearance is reality. So on this view, even envatted beings have true beliefs about the world.

I have recently found myself embracing a similar conclusion, though for quite different reasons. I don’t find the view that appearance is reality plausible, so I don’t endorse Berkeley’s reasoning. And until recently, it has seemed quite obvious to me that brains in vats would have massively false beliefs. But I now think there is a line of reasoning that shows that this is wrong.

I still think I cannot rule out the hypothesis that I am in a matrix. But I think that even if I am in a matrix, I am still in Tucson, I am still sitting at my desk, and so on. So the hypothesis that I am in a matrix is not a skeptical hypothesis. The same goes for Neo. At the beginning of the film, if he thinks “I have hair”, he is correct. If he thinks “It is sunny outside”, he is correct. And the same goes, of course, for the original brain in a vat. When it thinks “I have a body”, it is correct. When it thinks “I am walking”, it is correct.

This view may seem very counterintuitive at first. Initially, it seemed quite counterintuitive to me. So I’ll now present the line of reasoning that has convinced me that it is correct.

3 The Metaphysical Hypothesis

I will argue that the hypothesis that I am envatted is not a skeptical hypothesis, but a *metaphysical hypothesis*. That is, it is a hypothesis about the underlying nature of reality.

Where physics is concerned with the microscopic processes that underlie macroscopic reality, metaphysics is concerned with the fundamental nature of reality. A metaphysical hypothesis might make a claim about the reality that underlies physics itself. Alternatively, it might say something about the nature of our minds, or the creation of our world.
I think the Matrix Hypothesis should be regarded as a metaphysical hypothesis with all three of these elements. It makes a claim about the reality underlying physics, about the nature of our minds, and about the creation of the world.

In particular, I think the Matrix Hypothesis is equivalent to a version of the following three-part Metaphysical Hypothesis. First, physical processes are fundamentally computational. Second, our cognitive systems are separate from physical processes, but interact with these processes. Third, physical reality was created by beings outside physical space-time.

Importantly, nothing about this Metaphysical Hypothesis is skeptical. The Metaphysical Hypothesis here tells us about the processes underlying our ordinary reality, but it does not entail that this reality does not exist. We still have bodies, and there are still chairs and tables: it’s just that their fundamental nature is a bit different from what we may have thought. In this manner, the Metaphysical Hypothesis is analogous to a physical hypothesis, such as one involving quantum mechanics. Both the physical hypothesis and the Metaphysical Hypothesis tells us about the processes underlying chairs. They do not entail that there are no chairs. Rather, they tell us what chairs are really like. …

5 Life in the Matrix

The Matrix Hypothesis is not a skeptical hypothesis. If I accept it, I should not infer that the external world does not exist, or that I have no body, or that there are no tables and chairs, or that I am not in Tucson. Rather, I should infer that the physical world is constituted by computations beneath the microphysical level. There are still tables, chairs, and bodies: these are made up fundamentally of bits, and of whatever constitutes these bits. This world was created by other beings, but is still perfectly real. My mind is separate from physical processes, and interacts with them. My mind may not have been created by these beings, and it may not be made up of bits, but it still interacts with these bits.

The result is a complex picture of the fundamental nature of reality. The picture is strange and surprising, perhaps, but it is a picture of a full-blooded external world. If we are in a matrix, this is simply the way that the world is.

We can think of the Matrix Hypothesis as a creation myth for the information age. If it is correct, then the physical world was created, not necessarily by gods. Underlying the physical world is a giant computation, and creators created this world by implementing this computation. And our minds lie outside this physical structure, with an independent nature that interacts with this structure.

Many of the same issues that arise with standard creation myths arise here. When was the world created? Strictly speaking, it was not created within our time at all. When did history begin? The creators might have started the simulation in 4004 BC (or in 1999) with the fossil record intact, but it would have been much easier for them to start the simulation at the Big Bang and let things run their course from there. …

Even if we are not in a matrix, we can extend a version of this reasoning to other beings who are in a matrix. If they discover their situation, and come to accept that they are in a matrix, they should not reject their ordinary beliefs about the external world. At most, they should come to revise their beliefs about the underlying nature of their world: they should come to accept that external objects are made of bits, and so on. These beings are not massively deluded: most of their ordinary beliefs about their world are correct.
There are a few qualifications here. One may worry about beliefs about other people’s minds. I believe that my friends are conscious. If I am in a matrix, is this correct? In the Matrix depicted in the movie, these beliefs are mostly fine. This is a multi-vat matrix: for each of my perceived friends, there is an envatted being in the external reality, who is presumably conscious like me. The exception might be beings such as Agent Smith, who is not envatted, but is entirely computational. Whether these beings are conscious depends on whether computation is enough for consciousness. I will remain neutral on that issue here. We could circumvent this issue by building into the Matrix Hypothesis the requirement that all the beings we perceive are envatted. But even if we do not build in this requirement, we are not much worse off than in the actual world, where there is a legitimate issue about whether other beings are conscious, quite independently of whether we are in a matrix.

One might also worry about beliefs about the distant past, and about the far future. These will be unthreatened as long as the computer simulation covers all of space-time, from the Big Bang until the end of the universe. This is built into the Metaphysical Hypothesis, and we can stipulate that it is built into the Matrix Hypothesis too, by requiring that the computer simulation be a simulation of an entire world. There may be other simulations that start in the recent past (perhaps the Matrix in the movie is like this), and there may be others that only last for a short while. In these cases, the envatted beings will have false beliefs about the past and/or the future in their worlds. But as long as the simulation covers the lifespan of these beings, it is plausible that they will have mostly correct beliefs about the current state of their environment.

There may be some respects in which the beings in a matrix are deceived. It may be that the creators of the matrix control and interfere with much of what happens in the simulated world. (The Matrix in the movie may be like this, though the extent of the creators’ control is not quite clear.) If so, then these beings may have much less control over what happens than they think. But the same goes if there is an interfering god in a non-matrix world. And the Matrix Hypothesis does not imply that the creators interfere with the world, though it leaves the possibility open. At worst, the Matrix Hypothesis is no more skeptical in this respect than the Creation Hypothesis in a non-matrix world.

The inhabitants of a matrix may also be deceived in that reality is much bigger than they think. They might think their physical universe is all there is, when in fact there is much more in the world, including beings and objects that they can never possibly see. But again, this sort of worry can arise equally in a non-matrix world. For example, cosmologists seriously entertain the hypothesis that our universe may stem from a black hole in the “next universe up”, and that in reality there may be a whole tree of universes. If so, the world is also much bigger than we think, and there may be beings and objects that we can never possibly see. But either way, the world that we see is perfectly real.

Importantly, none of these sources of skepticism — about other minds, the past and the future, about our control over the world, and about the extent of the world — casts doubt on our belief in the reality of the world that we perceive. None of them leads us to doubt the existence of external objects such as tables and chairs, in the way that the vat hypothesis is supposed to do. And none of these worries is especially tied to the matrix scenario. One can raise doubts about whether other minds exist, whether the past and the future exist, and whether we have control over our worlds quite independently of whether we are in a matrix. If this is right, then the Matrix Hypothesis does not raise the distinctive skeptical issues that it is often taken to raise.
I suggested before that it is not out of the question that we really are in a matrix. One might have thought that this is a worrying conclusion. But if I am right, it is not nearly as worrying as one might have thought. Even if we are in such a matrix, our world is no less real than we thought it was. It just has a surprising fundamental nature. …

7 Other Objections

When we look at a brain in a vat from the outside, it is hard to avoid the sense that it is deluded. This sense manifests itself in a number of related objections. These are not direct objections to the argument above, but they are objections to its conclusion.

Objection 1: A brain in a vat may think it is outside walking in the sun, when in fact it is alone in a dark room. Surely it is deluded!

Response: The brain is alone in a dark room. But this does not imply that the person is alone in a dark room. By analogy, just say Descartes is right that we have disembodied minds outside space-time, made of ectoplasm. When I think “I am outside in the sun”, an angel might look at my ectoplasmic mind and note that in fact it is not exposed to any sun at all. Does it follow that my thought is incorrect? Presumably not: I can be outside in the sun, even if my ectoplasmic mind is not. The angel would be wrong to infer that I have an incorrect belief. Likewise, we should not infer that envatted being has an incorrect belief. At least, it is no more deluded than a Cartesian mind.

The moral is that the immediate surroundings of our minds may well be irrelevant to the truth of most of our beliefs. What matters is the processes that our minds are connected to, by perceptual inputs and motor outputs. Once we recognize this, the objection falls away.

Objection 2: An envatted being may believe that it is in Tucson, when in fact it is in New York, and has never been anywhere near Tucson. Surely this belief is deluded.

Response: The envatted being’s concept of “Tucson” does not refer to what we call Tucson. Rather, it refers to something else entirely: call this Tucson*, or “virtual Tucson”. We might think of this as a “virtual location” (more on this in a moment). When the being says to itself “I am in Tucson”, it really is thinking that it is in Tucson*, and it may well in fact be in Tucson*. Because Tucson is not Tucson*, the fact that the being has never been in Tucson is irrelevant to whether its belief is true.
A rough analogy: I look at my colleague Terry, and think “that’s Terry”. Elsewhere in the world, a duplicate of me looks at a duplicate of Terry. It thinks “that’s Terry”, but it is not looking at the real Terry. Is its belief false? It seems not: my duplicate’s “Terry” concept refers not to Terry, but to his duplicate Terry*. My duplicate really is looking at Terry*, so its belief is true. The same sort of thing is happening in the case above.

**Objection 3:** Before he leaves the Matrix, Neo believes that he has hair. But in reality he has no hair (the body in the vat is bald). Surely this belief is deluded.

Response: This case is like the last one. Neo’s concept of “hair” does not refer to real hair, but to something else that we might call hair* (“virtual hair”). So the fact that Neo does not have real hair is irrelevant to whether his belief is true. Neo really does have virtual hair, so he is correct. Likewise, when a child in the movie tells Neo “There is no spoon”, his concept refers to a virtual spoon, and there really is a virtual spoon. So the child is wrong.

**Objection 4:** What sort of objects does an envatted being refer to? What is virtual hair, virtual Tucson, and so on?

Response: These are all entities constituted by computational processes. If I am envatted, then the objects that I refer to (hair, Tucson, and so on) are all made of bits. And if another being is envatted, the objects that it refers to (hair*, Tucson*, and so on) are likewise made of bits. If the envatted being is hooked up to a simulation in my computer, then the objects it refers to are constituted by patterns of bits inside my computer. We might call these things *virtual objects*. Virtual hands are not hands (assuming I am not envatted), but they exist inside the computer all the same. Virtual Tucson is not Tucson, but it exists inside the computer all the same.

**Objection 5:** You just said that virtual hands are not real hands. Does this mean that if we are in the matrix, we don’t have real hands?

Response: No. If we are not in the matrix, but someone else is, we should say that their term “hand” refers to virtual hands, but our term does not. So in this case, our hands aren’t virtual hands. But if we are in the matrix, then our term “hand” refers to something that’s made of bits: virtual hands, or at least something that would be regarded as virtual hands by people in the next world up. That is, if we are in the matrix, real hands are made of bits. Things look quite different, and our words refer to different things, depending on whether our perspective is inside or outside the matrix.

This sort of perspective shift is common in thinking about the matrix scenario. From the first-person perspective, we suppose that we are in a matrix. Here, real things in our world are made of bits, though the “next world up” might not be made of bits. From the third-person perspective, we suppose that someone else is in a matrix but we are not. Here, real things in our world are not made of bits, but the “next world down” is made of bits. On the first way of doing things, our words refer to computational entities. On the second way of doing things, the envatted beings’ words refer to computational entities, but our words do not.

**Objection 6:** Just which pattern of bits is a given virtual object? Surely it will be impossible to pick out a precise set.

Response: This question is like asking: just which part of the quantum wavefunction is this chair, or is the University of Arizona? These objects are all ultimately constituted by an
underlying quantum wavefunction, but there may be no precise part of the micro-level wavefunction that we can say “is” the chair or the university. The chair and the university exist at a higher level. Likewise, if we are envatted, there may be no precise set of bits in the micro-level computational process that is the chair or the university. These exist at a higher level. And if someone else is envatted, there may be no precise sets of bits in the computer simulation that “are” the objects they refer to. But just as a chair exists without being any precise part of the wavefunction, a virtual chair may exist without being any precise set of bits.

**Objection 7:** An envatted being thinks it performs actions, and it thinks it has friends. Are these beliefs correct?

Response: One might try to say that the being performs actions* and that it has friends*. But for various reason I think it is not plausible that words like “action” and “friend” can shift their meanings as easily as words like like “Tucson” and “hair”. Instead, I think one can say truthfully (in our own language) that the envatted being performs actions, and that it has friends. To be sure, it performs actions in its environment, and its environment is not our environment but the virtual environment. And its friends likewise inhabit the virtual environment (assuming that we have a multi-vat matrix, or that computation suffices for consciousness). But the envatted being is not incorrect in this respect.

**Objection 8:** Set these technical points aside. Surely, if we are in a matrix, the world is nothing like we think it is!

Response: I deny this. Even if we are in a matrix, there are still people, football games, and particles, arranged in space-time just as we think they are. It is just that the world has a further nature that goes beyond our initial conception. In particular, things in the world are realized computationally in a way that we might not have originally imagined. But this does not contradict any of our ordinary beliefs. At most, it will contradict a few of our more abstract metaphysical beliefs. But exactly the same goes for quantum mechanics, relativity theory, and so on.

If we are in a matrix, we may not have many false beliefs, but there is much knowledge that we lack. For example, we do not know that the universe is realized computationally. But this is just what one should expect. Even if we are not in a matrix, there may well be much about the fundamental nature of reality that we do not know. We are not omniscient creatures, and our knowledge of the world is at best partial. This is simply the condition of a creature living in a world.

***PHIL 306 (Skepticism) students: Please continue reading. Otherwise: You may stop here.***

8 Other Skeptical Hypotheses

**New Matrix Hypothesis:** I was recently created, along with all my memories, and was put in a newly-created matrix.

What if both the matrix and I have existed for only a short time? This hypothesis is a computational version of Bertrand Russell’s Recent Creation Hypothesis: the physical world was created only recently (with fossil record intact), and so was I (with memories intact). On that hypothesis, the external world that I perceive really exists, and most of my beliefs about its current states are plausibly true, but I have many false beliefs about the past. I think the same should be said of the New Matrix Hypothesis. One can argue, along the lines presented
earlier, that the New Matrix Hypothesis is equivalent to a combination of the Metaphysical Hypothesis with the Recent Creation Hypothesis. This combination is not a global skeptical hypothesis (though it is a partial skeptical hypothesis, where beliefs about the past are concerned). So the same goes for the New Matrix Hypothesis.

**Recent Matrix Hypothesis:** For most of my life I have not been envatted, but I was recently hooked up to a matrix.

If I was recently put in a matrix without realizing it, it seems that many of my beliefs about my current environment are false. Let’s say that just yesterday someone put me into a simulation, in which I fly to Las Vegas and gamble at a casino. Then I may believe that I am in Las Vegas now, and that I am in a casino, but these beliefs at false: I am really in a laboratory in Tucson.

This result is quite different from the long-term matrix. The difference lies in the fact that my conception of external reality is anchored to the reality in which I have lived most of my life. If I have been envatted all my life, my conception is anchored to the computationally constituted reality. But if I was just envatted yesterday, my conception is anchored to the external reality. So when I think that I am in Las Vegas, I am thinking that I am in the external Las Vegas, and this thought is false.

Still, this does not undercut all of my beliefs about the external world. I believe that I was born in Sydney, that there is water in the oceans, and so on, and all of these beliefs are correct. It is only my recently acquired beliefs, stemming from perception of the simulated environment, that will be false. So this is only a partial skeptical hypothesis: its possibility casts doubt on a subset of our empirical beliefs, but it does not cast doubt on all of them.

Interestingly, the Recent Matrix and the New Matrix hypothesis give opposite results, despite their similar nature: the Recent Matrix Hypothesis yields true beliefs about the past but false beliefs about the present, while the New Matrix Hypothesis yields false beliefs about the past and true beliefs about the present. The differences are tied to the fact that in Recent Matrix Hypothesis, I really have a past existence for my beliefs to be about, and that past reality has played a role in anchoring the contents of my thoughts that has no parallel under the New Matrix Hypothesis.

**Local Matrix Hypothesis:** I am hooked up to a computer simulation of a fixed local environment in a world.

On one way of doing this, a computer simulates a small fixed environment in a world, and the subjects in the simulation encounter some sort of barrier when they try to leave that area. For example, in the movie *The Thirteenth Floor*, just California is simulated, and when the subject tries to drive to Nevada, the road says “Closed for Repair” (with faint green electronic mountains in the distance!). Of course this is not the best way to create a matrix, as subjects are likely to discover the limits to their world.

This hypothesis is analogous to a Local Creation Hypothesis, on which creators just created a local part of the physical world. Under this hypothesis, we will have true beliefs about nearby matters, but false beliefs about matters further from home. By the usual sort of reasoning, the Local Matrix Hypothesis can be seen as a combination of the Metaphysical Hypothesis with the Local Creation Hypothesis. So we should say the same thing about this.
**Extendible Local Matrix Hypothesis:** I am hooked up to a computer simulation of a local environment in a world, extended when necessary depending on subject’s movements.

This hypothesis avoids the obvious difficulties with a fixed local matrix. Here the creators simulate a local environment and extend it when necessary. For example, they might right now be concentrating on simulating a room in my house in Tucson. If I walk into another room, or fly to another city, they will simulate those. Of course they need to make sure that when I go to these places, they match my memories and beliefs reasonably well, with allowance for evolution in the meantime. The same goes for when I encounter familiar people, or people I have only heard about. Presumably the simulators keep up a database of the information about the world that has been settled so far, updating this information whenever necessary as time goes along, and making up new details when they need them.

This sort of simulation is quite unlike simulation in an ordinary matrix. In a matrix, the whole world is simulated at once. There are high start-up costs, but once the simulation is up and running, it will take care of itself. By contrast, the extendible local matrix involves “just-in-time” simulation. This has much lower start-up costs, but it requires much more work and creativity as the simulation evolves.

This hypothesis is analogous to an Extendible Local Creation Hypothesis about ordinary reality, under which creators create just a local physical environment, and extend it when necessary. Here, external reality exists and many local beliefs are true, but again beliefs about matters further from home are false. If we combine that hypothesis with the Metaphysical Hypothesis, the result is the Extendible Local Matrix Hypothesis. So if we are in an extendible local matrix, external reality still exists, but there is not as much of it as we thought. Of course if I travel in the right direction, more of it may come into existence!

The situation is reminiscent of *The Truman Show*. Truman lives in an artificial environment made up of actors and props, which behave appropriately when he is around, but which may be completely different when he is absent. Truman has many true beliefs about his current environment: there really are tables and chairs in front of him, and so on. But he is deeply mistaken about things outside his current environment, and further from home.

It is common to think that while *The Truman Show* poses a disturbing skeptical scenario, *The Matrix* is much worse. But if I am right, things are reversed. If I am in a matrix, then most of my beliefs about the external world are true. If I am in something like *The Truman Show*, then a great number of my beliefs are false. On reflection, it seems to me that this is the right conclusion. If we were to discover that we were (and always had been) in a matrix, this would be surprising, but we would quickly get used to it. If we were to discover that we were (and always had been) in the Truman Show, we might well go insane.

**Macroscopic Matrix Hypothesis:** I am hooked up to a computer simulation of macroscopic physical processes without microphysical detail.

One can imagine that for ease of simulation, the makers of a matrix might not both to simulate low-level physics. Instead, they might just represent macroscopic objects in the world and their properties: e.g. that there is a table with such-and-such shape, position, and color, with a book on top of it with certain properties, and so on. They will need to make some effort to
make sure that these objects behave in a physically reasonable way, and they will have to make special provisions for handling microphysical measurements, but one can imagine that at least a reasonable simulation could be created this way.

I think this hypothesis is analogous to a Macroscopic World Hypothesis: there are no microphysical processes, and instead macroscopic physical objects exist as fundamental objects in the world, with properties of shape, color, position, and so on. This is a coherent way our world could be, and it is not a global skeptical hypothesis, though it may lead to false scientific beliefs about lower levels of reality. The Macroscopic Matrix Hypothesis can be seen as a combination of this hypothesis with a version of the Metaphysical Hypothesis. As such, it is not a global skeptical hypothesis either.

One can also combine the various hypothesis above in various ways, yielding hypotheses such as a New Local Macroscopic Matrix Hypothesis. For the usual reasons, all of these can be seen as analogs of corresponding hypotheses about the physical world. So all of them are compatible with the existence of physical reality, and none is a global skeptical hypothesis.

God Hypothesis: Physical reality is represented in the mind of God, and our own thoughts and perceptions depend on God’s mind.

A hypothesis like this was put forward by George Berkeley as a view about how our world might really be. Berkeley intended this as a sort of metaphysical hypothesis about the nature of reality. Most other philosophers have differed from Berkeley in regarding this as a sort of skeptical hypothesis. If I am right, Berkeley is closer to the truth. The God Hypothesis can be seen as a version of the Matrix Hypothesis, on which the simulation of the world is implemented in the mind of God. If this is right, we should say that physical processes really exist: it’s just that at the most fundamental level, they are constituted by processes in the mind of God.

Evil Genius Hypothesis: I have a disembodied mind, and an evil genius is feeding me sensory inputs to give the appearance of an external world.

This is Rene Descartes’s classical skeptical hypothesis. What should we say about it? This depends on just how the evil genius works. If the evil genius simulates an entire world in his head in order to determine what inputs I should receive, then we have a version of the God Hypothesis. Here we should say that physical reality exists and is constituted by processes within the genius. If the evil genius is simulating only a small part of the physical world, just enough to give me reasonably consistent inputs, then we have an analog of the Local Matrix Hypothesis (in either its fixed or flexible versions). Here we should say that just a local part of external reality exists. If the evil genius is not bothering to simulate the microphysical level, but just the macroscopic level, then we have an analog of the Macroscopic Matrix Hypothesis. Here we should say that local external macroscopic objects exist, but our beliefs about their microphysical nature are incorrect.

The evil genius hypothesis is often taken to be a global skeptical hypothesis. But if the reasoning above is right, this is incorrect. Even if the Evil Genius Hypothesis is correct, some of the external reality that we apparently perceive really exists, though we may have some false beliefs about it, depending on details. It is just that this external reality has an underlying nature that is quite different from what we may have thought.
**Dream Hypothesis:** I am now and have always been dreaming.

Descartes raised the question: how do you know that you are not currently dreaming? Morpheus raises a similar question:

Have you ever had a dream, Neo, that you were so sure was real. What if you were unable to wake from that dream? How would you know the difference between the dream world and the real world?

The hypothesis that I am currently dreaming is analogous to a version of the Recent Matrix Hypothesis. I cannot rule it out conclusively, and if it is correct, then many of my beliefs about my current environment are incorrect. But presumably I still have many true beliefs about the external world, anchored in the past.

What if I have always been dreaming? That is, what if all of my apparent perceptual inputs have been generated by my own cognitive system, without my realizing this? I think this case is analogous to the Evil Genius Hypothesis: it’s just that the role of the “evil genius” is played by a part of my own cognitive system! If my dream-generating system simulates all of space-time, we have something like the original Matrix Hypothesis. If it models just my local environment, or just some macroscopic processes, we have analogs of the more local versions of the Evil Genius Hypothesis above. In any of these cases, we should say that the objects that I am currently perceiving really exist (although objects farther from home may not). It is just that some of them are constituted by my own cognitive processes.

**Chaos Hypothesis:** I do not receive inputs from anywhere in the world. Instead, I have random uncaused experiences. Through a huge coincidence, they are exactly the sort of regular, structured experiences with which I am familiar.

The Chaos Hypothesis is an extraordinarily unlikely hypothesis, much more unlikely than anything considered above. But it is still one that could in principle obtain, even if it has minuscule probability. If I am chaotically envatted, do physical processes in the external world exist? I think we should say that they do not. My experiences of external objects are caused by nothing, and the set of experiences associated with my conception of a given object will have no common source. Indeed, my experiences are not caused by any reality external to them at all. So this is a genuine skeptical hypothesis: if accepted, it would cause us to reject most of our beliefs about the external world.

So far, the only clear case of a global skeptical hypothesis is the Chaos Hypothesis. Unlike the previous hypothesis, accepting this hypothesis would undercut all of our substantive beliefs about the external world. Where does the difference come from?

Arguably, what is crucial is that on the Chaos Hypothesis, there is no causal explanation of our experiences at all, and there is no explanation for the regularities in our experience. In all the previous cases, there is some explanation for these regularities, though perhaps not the explanation that we expect. One might suggest that as long as a hypothesis involves some reasonable explanation for the regularities in our experience, then it will not be a global skeptical hypothesis.
If so, then if we are granted the assumption that there is some explanation for the regularities in our experience, then it is safe to say that some of our beliefs about the external world are correct. This is not much, but it is something!

9 Philosophical Notes

The material above was written to be accessible to a wide audience, so it deliberately omits technical philosophical details, connections to the literature, and so on. Here I will try to remedy this omission. Readers without a background in philosophy should probably skip or skim this section.

Note 1: Hilary Putnam (1981) has argued that the hypothesis that I am (and have always been) a brain in a vat can be ruled out a priori. In effect, this is because my word “brain” refers to objects in my perceived world, and it cannot refer to objects in an “outer” world in which the vat would have to exist. For my hypothesis “I am a brain in a vat” to be true, I would have to be a brain of the sort that exists in the perceived world, but that cannot be the case. So the hypothesis must be false.

An analogy: I can arguably rule out the hypothesis that I am in the Matrix (capital M). My term “the Matrix” refers to a specific system that I have seen in a movie in my perceived world. I could not be in that very system, as the system exists within the world that I perceive. So my hypothesis “I am in the Matrix” must be false.

This conclusion about the Matrix seems reasonable, but there is a natural response. Perhaps this argument rules out the hypothesis that I am in the Matrix, but I cannot rule out the hypothesis that I am in a matrix, where a matrix is a generic term for a computer simulation of a world. The term “Matrix” may be anchored to the specific system in the movie, but the generic term “matrix” is not.

Likewise, it is arguable that I can rule out the hypothesis that I am a brain in a vat (if “brain” is anchored to a specific sort of biological system in my perceived world). But I cannot rule out the hypothesis that I am envatted, where I have a cognitive system that receives input from and sends outputs to a computer simulation of a world. The term “envatted” (and the terms used in its definition) are generic terms, not anchored to specific systems in perceived reality. By using this slightly different language, we can restate the skeptical hypothesis in a way that is invulnerable to Putnam’s reasoning.

More technically: Putnam’s argument may work for “brain” and “Matrix” because one is a natural kind term and the other is a name. These terms are subject to “Twin Earth” thought experiments (Putnam 1975), where duplicates can use corresponding terms with different referents. On Earth, Oscar’s term “water” refers to H₂O; but on Twin Earth (which contains the superficially identical XYZ in its oceans and lakes), Twin Oscar’s term “water” refers to XYZ. Likewise, perhaps my term “brain” refers to biological brains, while an envatted being’s term “brain” refers to virtual brains. If so, when an envatted being says “I am a brain in a vat”, it is not referring to its biological brain, and its claim is false.

But not all terms are subject to Twin Earth thought experiments. In particular, semantically neutral terms are not (at least when used without semantic deference): such terms plausibly include “philosopher”, “friend”, and many others. Other such terms include “matrix” and “envatted”, as defined in this article. If we work with hypotheses such as “I am in a matrix”
and “I am envatted”, rather than “I am in the Matrix” or “I am a brain in a vat”, then Putnam’s argument does not apply. Even if a brain in a vat could not truly think “I am a brain in a vat”, it could truly think “I am envatted”. So I think that Putnam’s line of reasoning is ultimately a red herring.

**Note 2:** Despite this disagreement, the conclusion of this article is closely related to another suggestion of Putnam’s. This is the suggestion that a brain in a vat may have true beliefs, because it will refer to chemical processes or processes inside a computer. However, I reach this conclusion by a quite different route. Putnam argues by an appeal to the causal theory of reference: thoughts refer to what they are causally connected to, and the thoughts of an envatted being are causally connected to processes in a computer. This argument is clearly inconclusive, as the causal theory of reference is so unconstrained. To say that a causal connection is required for reference is not to say what sort of causal connection suffices. There are many cases (like “phlogiston”) where terms fail to refer despite rich causal connections. Intuitively, it is natural to think that the brain in a vat is a case like this, so an appeal to the causal theory of reference does not seem to help.

The argument I have given presupposes nothing about the theory of reference. Instead, it proceeds directly by considering first-order hypotheses about the world, the connections between these, and what we should say if they are true. In answering objections, I have made some claims about reference, and these claims are broadly compatible with a causal theory of reference. But importantly, these claims are very much consequences of the first-order argument rather than presuppositions of it. In general, I think that claims in the theory of reference are beholden to first-order judgments about cases, rather than vice versa. …

**Note 15:** The reasoning in this paper does not offer a knockdown refutation of skepticism, as several skeptical hypotheses are left open. But I think it significantly strengthens one of the standard responses to skepticism. It is often held that although various skeptical hypotheses are compatible with our experiences, the hypothesis that there is a real physical world provides a simpler or better explanation of the regularities in our experiences than these skeptical hypotheses. If so, then we may be justified in believing in the real physical world, by an inference to the best explanation.

At this point is often objected that some skeptical hypotheses seem just as simple as the standard explanation: for example, the hypothesis that all our experiences are caused by a computer simulation, or by God. If so, this response to skepticism fails. But if I am right, then these “equally simple” hypotheses are not skeptical hypotheses at all. If so, then inference to the best explanation may work after all: all of these “simple” hypotheses yield mostly true beliefs about an external world.

The residual issue concerns the various remaining skeptical hypotheses on the table, such as the Recent Matrix Hypothesis, the Local Matrix Hypothesis, and so on. It seems reasonable to hold that these are significantly less simple than the hypotheses above, however. All of them involve a non-uniform explanation of the regularities in our experiences. In the Recent Matrix Hypothesis, present regularities and past regularities have very different explanations. In the Local Matrix Hypothesis, beliefs about matters close to home and far from home have very different explanations. These hypotheses as a whole have a sort of dual-mechanism structure that seems considerably more complex than the uniform-mechanism structures above. If this is right, one can argue that inference to the best explanation justifies us in ruling out these hypotheses, and in accepting the non-skeptical hypotheses above.
Even if one thinks that some of these skeptical hypotheses offer reasonably good explanations of our experience, there is still a promising argument against global external-world skepticism in the vicinity. If I am right, all of these skeptical hypotheses are at worst partial skeptical hypotheses: if they are correct, then a good many of our empirical beliefs will still be true, and there will still be an external world. To obtain a global skeptical hypothesis, we have to go all the way to the Chaos Hypothesis. But this is a hypothesis on which the regularities in our experience have no explanation at all. Even an extremely weak version of inference to the best explanation justifies us in ruling out this sort of hypothesis. If so, then this sort of reasoning may justify our belief in the existence of the external world.

References


