PHILOSOPHY & THE MATRIX

Click Below:

November 20, 2002
Launch Announcement

March 20, 2003
Update Announcement

News / Introduction

The Matrix is a film that astounds not only with action and special effects but also with ideas. These pages are dedicated to exploring some of the many philosophical ideas that arise in both the original film and the sequels. In the upcoming months we will be continually expanding this section, offering essays from some of the brightest minds in philosophy and cognitive science. News about updates to this section can be found right here. *(Scroll down to read about the latest update, or just click here.)*

LAUNCH: NOVEMBER 20, 2002

We are kicking things off with essays from eight different contributors on various philosophical, technological, and religious aspects of the film.

Though this collection of essays is part of the official web site for the Matrix films, the views expressed in these essays are *solely those of the individual authors*. The Wachowski brothers have remained relatively tight-lipped regarding the religious symbolism and philosophical themes that permeate the film, preferring that the movie speak for itself. Accordingly, you will not find anyone here claiming to offer *the* definitive analysis of the film, its symbols, message, etc. What you will find instead are essays that both elucidate the philosophical problems raised by the film and explore possible avenues for solving these problems. Some of these essays are more pedagogical in nature – instructing the reader in the various ways in which *The Matrix* raises questions that have been tackled throughout history by prominent philosophers. Other contributors use the film as a springboard for discussing their *own* original philosophical views. As you will see, the authors don't always agree with each other regarding how best to interpret the film. However, all of

the essays share the aim of giving the reader a sense of how this remarkable film offers more than the standard Hollywood fare. In other words, their common goal is to help show you just "how deep the rabbit-hole goes."

Beginning the collection are three short essays in which I discuss two of the more conspicuous philosophical questions raised by the film: the skeptical worry that one's experience may be illusory, and the moral question of whether it matters. Highlighting the parallels between the scenario described in *The Matrix* and similar imaginary situations that have been much discussed by philosophers, these essays offer an introduction to the positions taken by various thinkers on these fascinating skeptical and moral puzzles. They serve as a warm-up for things to come.

Next is "The Matrix of Dreams" by **Colin McGinn**, a distinguished contemporary philosopher who is perhaps best known for his writings on consciousness. His essay offers an analysis of the film that focuses on the dreamlike nature of the world of the Matrix. Arguing that it is misguided to characterize the situation described by the film as involving hallucinations, McGinn seeks to show how the particular details of the film make it more plausible to see the Matrix as involving the direct employment of one's *imagination* (as in a dream), rather than a force-feeding of false *perceptions*. Along the way, McGinn's essay also touches on the moral assumptions of the film, several other philosophical problems raised by the character of Cypher, and the dreamlike quality of *all* films.

Hubert Dreyfus is a philosopher known both for his pioneering discussion of the philosophical problems of Artificial Intelligence, and his work in bridging the gap between recent European and English-language philosophy. In "The Brave

New World of *The Matrix*," he and his son **Stephen Dreyfus** draw on the phenomenological tradition that began with Edmund Husserl and culminates in Maurice Merleau-Ponty to discuss the skeptical and moral problems raised by the film. They argue that the real worry facing folks trapped in the Matrix involves not deception or the possession of possibly false beliefs, but the limits on creativity imposed by the Matrix. Following Martin Heidegger in suggesting that our human nature lies in our capacity to redefine our nature and thereby open up new worlds, they conclude that this capacity for radical creation seems unavailable to those locked within the pre-programmed confines of the Matrix.

Richard Hanley, author of the best-selling book *The Metaphysics of Star Trek* and a philosophy professor at the University of Delaware, again explores the intersection of philosophy and science fiction with his entertaining and thought-provoking piece "Never the Twain Shall Meet: Reflections on The First Matrix." In it he argues that *The Matrix* may have lessons to teach us regarding the coherence of our values. In particular, he makes the case that, given a traditional Christian notion of an afterlife, Heaven turns out to be rather like a Matrix! Even more surprising is a corollary to this thesis: Jean-Paul ("Hell is other people") Sartre was close to the truth after all – Heaven is best understood as a Matrix-like simulation in which contact with other real human beings is eliminated.

Iakovos Vasiliou, a philosopher at Brooklyn College who specializes in Plato, Aristotle, and Wittgenstein, offers a penetrating investigation into the differences (and surprising similarities) between the scenario described in *The Matrix* and our own everyday situation in his essay "Reality, What Matters, and *The Matrix*." Pointing out that more than we might expect hinges on the moral backdrop of *The Matrix* plot line, he asks readers to instead envisage a "benevolently generated Matrix." Given the possibility of such a Matrix and the

actuality of a horrible situation on Earth, he argues that we will agree that entering into it offers not a denial of what we most value but instead a chance to better realize those values.

Changing gears a bit we then have an essay from the notable (and some would say notorious) cybernetics pioneer **Kevin Warwick**. He is known internationally for his robotics research and in particular for a series of procedures in which he was implanted with sensors that connected him to computers and the internet. Less well-publicized is the fact that several years before *The Matrix* came out he published a non-fiction book that predicted the ultimate takeover of mankind by a race of super-intelligent robots. In his contribution here ("*The Matrix* – Our Future?") he draws on his years of research to muse on the plausibility (and desirability) of the scenario described in *The Matrix*, concluding that a real-life Matrix need not be feared if we prepare ourselves adequately. How? By becoming part machine ourselves – Warwick argues that transforming ourselves into Cyborgs will allow us to "plug in" confident that we will fully benefit from all that such a future offers.

Rounding out our collection is an essay entitled "Wake Up! Gnosticism & Buddhism in *The Matrix*" from two professors of religion: **Frances Flannery-Dailey** and **Rachel Wagner**. Flannery-Dailey's research speciality is ancient dreams, apocalypticism and early-Jewish mysticism, while Wagner's research focuses on biblical studies and the relationship between religion & culture. Their essay offers a comprehensive treatment of the Gnostic and Buddhist themes that appear in the film. While pointing out the many differences between these two traditions and the eclectic manner in which both are referenced throughout the film, Flannery-Dailey and Wagner make it clear that common to Gnosticism, Buddhism, and *The Matrix* is the idea that what we

take to be reality is in fact a kind of illusion or dream from which we ought best to "wake up." Only then can enlightenment, be it spiritual or otherwise, occur.

We hope you enjoy this first batch of essays. Check back for future contributions from the renowned philosopher of mind **David Chalmers** (Arizona), moral philosopher **Julia Driver** (Dartmouth), and epistemologist **James Pryor** (Princeton), among others.

UPDATE: MARCH 20, 2003

As promised, we are pleased to offer five new essays tackling philosophical themes that arise in *The Matrix*.

Starting things off is a piece by the epistemologist and philosopher of mind **James Pryor**. He's contributed a lively essay that will be of particular interest to those coming to philosophy for the first time. In "What's So Bad About Living in The Matrix?" he explores and criticizes two tempting but problematic philosophical positions: the view that there can't be facts which it's impossible for us to know about (sometimes called verificationism), and the view that everyone's motive for acting is always to have nicer experiences. Employing examples from both the film and imaginary thought-experiments, Pryor tries to show that these positions, which can often initially seem irresistible to students, are not as straightforward or as satisfying as they might first appear. He then goes on to argue (in sympathy with Vasiliou's essay) that the worst thing about living in the Matrix would not be the metaphysical or epistemological limitations such a scenario would impose, it would instead be the political constraints: those trapped in the Matrix have constraints on their

action that most of us deeply value not having.

David Chalmers is a philosopher from the University of Arizona and author of numerous books and articles on the philosophy of mind, including the influential volume *The Conscious Mind*. In his essay "The Matrix as Metaphysics," he suggests that while we cannot rule out the possibility that we are in a system like the Matrix, this possibility is not as bad as we might have thought. He argues against the intuitive view that if we are in a matrix, we are deluded about the external world. Instead, he suggests that if we are in a matrix, we should regard this as telling us about the nature of the external world: the physical world is ultimately made of bits, and was created by beings who ensured that our minds interact with this physical world. Chalmers's surprising conclusion is that even if we are living in a Matrix-like simulation, most of our beliefs about the world are still true.

Julia Driver, a moral philosopher from Dartmouth College and author of *Uneasy Virtue*, explores some of the distinctively ethical issues that arise in *The Matrix* in her essay "Artificial Ethics." Driver begins by using the film to consider the moral status of artificially created beings: she argues that, given certain assumptions regarding the nature of consciousness, rationality, and personhood, we ought to regard artificial intelligences such as Agent Smith as creatures that deserve genuine moral consideration. In the second part of her essay Driver tackles the thorny philosophical question of whether one can behave immorally when in “non-veridical” (illusory) circumstances. Noting the implausibility of attributing wrongdoing to those who perform seemingly immoral acts in a dream, she argues that, to the extent that the Matrix offers a similarly illusory world free of actual unpleasant effects on others, it also seems odd to attribute wrongdoing to agents acting in such a world. However,

drawing on insights from the first part of her essay, Driver concludes that we have good reasons to think that actions in the Matrix *would* have genuine effects on both humans and some artificial creatures, and thus the world of the Matrix, like our world, has its own moral norms — its own ethics — that ought to be both acknowledged and respected.

Michael McKenna, a philosopher at Ithaca College who specializes in the philosophical problems of freedom and moral responsibility, offers up a comprehensive yet light-hearted exploration of the free will problem in his essay "Neo's Freedom ... Whoa!". Ingeniously utilizing aspects of *The Matrix* to describe and explore the traditional positions taken in debates over free will, McKenna manages to cover a lot of ground: determinism, fatalism, compatibilism, and incompatibilism are all canvassed and compared through the unique perspective afforded us by the film. He then goes on to explore the attractiveness of the radical freedom that Neo appears to have achieved by the end of *The Matrix*. Does such absolute freedom indeed “rock” the way we naturally think it would? McKenna convincingly argues that total freedom of this sort offers too much of a good thing: part of the joy we take in exercising our freedom is in pushing boundaries and testing limits — if all boundaries and limitations are removed, the possibility for such joy will disappear as well.

Finally, we have an essay from **John Partridge**, a professor of philosophy at Wheaton College whose work focuses on the philosophy of the ancient Greeks. In "Plato's Cave & *The Matrix*," Partridge considers the striking similarities between *The Matrix* and the "cave" scenario described in Plato’s *Republic*. In addition to pointing out the numerous surface parallels between the cave-dwellers Plato describes and the humans trapped in the Matrix, Partridge explores a deeper continuity between the film and Plato’s text: both narratives

privilege the self-knowledge that follows from the right kind of self-examination. As Plato might put it, both Neo and the cave-dwellers must undertake a difficult journey from darkness to light if genuine knowledge (and consequently true "care of the soul") is to be attained.

Enjoy this new group of essays, and be sure and check back soon for further updates.

Chris Grau, Editor

Dream Skepticism

MORPHEUS:

Have you ever had a dream, Neo, that you were so sure was real?

MORPHEUS:

What if you were unable to wake from that dream, Neo? How would you know the difference between the dreamworld and the real world?

Neo has woken up from a hell of a dream — the dream that was his life. How was he to know? The cliché is that if you are dreaming and you pinch yourself, you will wake up. Unfortunately, things aren't quite that simple. It is the nature of most dreams that we take them for reality — while dreaming we are unaware that we are in fact in a dreamworld. Of course, we eventually wake up, and when we do we realize that our experience was all in our mind. Neo's predicament makes one wonder, though: how can any of us be sure that we have ever *genuinely* woken up? Perhaps, like Neo prior to his downing the red pill, our dreams thus far have in fact been dreams *within* a dream.

The idea that what we take to be the real world could all be just a dream is familiar to many students of philosophy, poetry, and literature. Most of us, at one time or another, have been struck with the thought that we might mistake a dream for reality, or reality for a dream. Arguably the most famous exponent of this worry in the Western philosophical tradition is the seventeenth-century French philosopher Rene Descartes. In an attempt to provide a firm foundation for knowledge, he began his *Meditations* by clearing the philosophical ground through doubting all that could be doubted. This was done, in part, in order to

determine if anything that could count as certain knowledge could survive such rigorous and systematic skepticism. Descartes takes the first step towards this goal by raising (through his fictional narrator) the possibility that we might be dreaming:

> "How often, asleep at night, am I convinced of just such familiar events — that I am here in my dressing gown, sitting by the fire —when in fact I am lying undressed in bed! Yet at the moment my eyes are certainly wide awake when I look at this piece of paper; I shake my head and it is not asleep; as I stretch out and feel my hand I do so deliberately, and I know what I am doing. All this would not happen with such distinctness to someone asleep. Indeed! As if I did not remember other occasions when I have been tricked by exactly similar thoughts while asleep! As I think about this more carefully, I see plainly that there are never any sure signs by means of which being awake can be distinguished from being asleep. The result is that I begin to feel dazed, and this very feeling only reinforces the notion that I may be asleep." (*Meditations*, 13)

When we dream we are often blissfully ignorant that we are dreaming. Given this, and the fact that dreams often seem as vivid and "realistic" as real life, how can you rule out the possibility that you might be dreaming even now, as you sit at your computer and read this? This is the kind of perplexing thought Descartes forces us to confront. It seems we have no justification for the belief that we are not dreaming. If so, then it seems we similarly have no justification in thinking that the world we experience is the real world. Indeed, it becomes questionable whether we are justified in thinking that *any* of our beliefs are true.

The narrator of Descartes' *Meditations* worries about this, but he ultimately maintains that the possibility that one might be dreaming cannot by itself cast doubt on all we think we know; he points out that even if all our sensory experience is but a dream, we can still conclude that we have *some* knowledge of the nature of reality. Just as a painter cannot create *ex nihilo* but must rely

on pigments with which to create her image, certain elements of our thought must exist prior to our imaginings. Among the items of knowledge that Descartes thought survived dream skepticism are truths arrived at through the use of reason, such as the truths of mathematics: "For whether I am awake or asleep, two and three added together are five, and a square has no more than four sides." (14)

While such an insight offers little comfort to someone wondering whether the people and objects she confronts are genuine, it served Descartes' larger philosophical project: he sought, among other things, to provide a foundation for knowledge in which truths arrived at through reason are given priority over knowledge gained from the senses. (This bias shouldn't surprise those who remember that Descartes was a brilliant mathematician in addition to being a philosopher.) Descartes was not himself a skeptic — he employs this skeptical argument so as to help remind the reader that the truths of mathematics (and other truths of reason) are on firmer ground than the data provided to us by our senses.

Despite the fact that Descartes' ultimate goal was to demonstrate how genuine knowledge is possible, he proceeds in *The Meditations* to utilize a much more radical skeptical argument, one that casts doubt on even his beloved mathematical truths. In the next section we will see that, many years before the Wachowskis dreamed up *The Matrix,* Descartes had imagined an equally terrifying possibility.

Further Reading:

Dancy, Jonathan. *Introduction to Contemporary Epistemology*, Blackwell, 1985.

Descartes.*The Philosophical Writings of Descartes*, tr: John Cottingham, Robert Stoothoff, Dugald Murdoch. Cambridge University Press, 1984

Stroud, Barry.*The Significance of Philosophical Scepticism*, Oxford, 1984.

NEXT ESSAY

Brains in Vats and The Evil Demon

MORPHEUS:

What is the Matrix? Control.

MORPHEUS:

The Matrix is a computer-generated dreamworld built to keep us under control in order to change a human being into this. (holds up a coppertop battery)

NEO:

No! I don't believe it! It's not possible!

Before breaking out of the Matrix, Neo's life was not what he thought it was. It was a lie. Morpheus described it as a "dreamworld," but unlike a dream, this world was not the creation of Neo's mind. The truth is more sinister: the world was a creation of the artificially intelligent computers that have taken over the Earth and have subjugated mankind in the process. These creatures have fed Neo a simulation that he couldn't possibly help but take as the real thing. What's worse, it isn't clear how any of us can know with certainty that we are not in a position similar to Neo before his "rebirth." Our ordinary confidence in our ability to reason and our natural tendency to trust the deliverances of our senses can both come to seem rather naive once we confront this possibility of deception.

A viewer of *The Matrix* is naturally led to wonder: how do I know I am not in the Matrix? How do I know for sure that my world is not also a sophisticated charade, put forward by some super-human intelligence in such a way that I cannot possibly detect the ruse? The philosopher Rene Descartes suggested a

similar worry: the frightening possibility that all of one's experiences might be the result of a powerful outside force, a "malicious demon."

> "And yet firmly implanted in my mind is the long-standing opinion that there is an omnipotent God who made me the kind of creature that I am. How do I know that he has not brought it about that there is no earth, no sky, no extended thing, no shape, no size, no place, while at the same time ensuring that all these things appear to me to exist just as they do now? What is more, just as I consider that others sometimes go astray in cases where they think they have the most perfect knowledge, how do I know that God has not brought it about that I too go wrong every time I add two and three or count the sides of a square, or in some even simpler matter, if that is imaginable? But perhaps God would not have allowed me to be deceived in this way, since he is said to be supremely good; [...] I will suppose therefore that not God, who is supremely good and the source of truth, but rather some malicious demon of the utmost power and cunning has employed all his energies in order to deceive me. I shall think that the sky, the air, the earth, colours, shapes, sounds and all external things are merely the delusions of dreams which he has devised to ensnare my judgment." (*Meditations,* 15)

The narrator of Descartes' *Meditations* concludes that none of his former opinions are safe. Such a demon could not only deceive him about his perceptions, it could conceivably cause him to go wrong when performing even the simplest acts of reasoning.

This radical worry seems inescapable. How could you possibly prove to yourself that you are not in the kind of nightmarish situation Descartes describes? It would seem that any argument, evidence or proof you might put forward could easily be yet another trick played by the demon. As ludicrous as the idea of the evil demon may sound at first, it is hard, upon reflection, not to share Descartes' worry: for all you know, you may well be a mere plaything of such a malevolent intelligence. More to the point of our general discussion: for all you know, you may well be trapped in the Matrix.

Many contemporary philosophers have discussed a similar skeptical dilemma

that is a bit closer to the scenario described in *The Matrix*. It has come to be known as the "brain in a vat" hypothesis, and one powerful formulation of the idea is presented by the philosopher Jonathan Dancy:

> "You do not know that you are not a brain, suspended in a vat full of liquid in a laboratory, and wired to a computer which is feeding you your current experiences under the control of some ingenious technician scientist (benevolent or malevolent according to taste). For if you were such a brain, then, provided that the scientist is successful, nothing in your experience could possibly reveal that you were; for your experience is *ex hypothesi* identical with that of something which is not a brain in a vat. Since you have only your own experience to appeal to, and that experience is the same in either situation, nothing can reveal to you which situation is the actual one." (*Introduction to Contemporary Epistemology*, 10)

If you cannot know whether you are in the real world or in the word of a computer simulation, you cannot be sure that your beliefs about the world are true. And, what was even more frightening to Descartes, in this kind of scenario it seems that your ability to reason is no safer than the deliverances of the senses: the evil demon or malicious scientist could be ensuring that your reasoning is just as flawed as your perceptions.

As you have probably already guessed, there is no easy way out of this philosophical problem (or at least there is no easy *philosophical* way out!). Philosophers have proposed a dizzying variety of "solutions" to this kind of skepticism but, as with many philosophical problems, there is nothing close to unanimous agreement regarding how the puzzle should be solved.

Descartes' own way out of his evil demon skepticism was to first argue that one cannot genuinely doubt the existence of oneself. He pointed out that all thinking presupposes a thinker: even in doubting, you realize that there must

at least be a self which is doing the doubting. (Thus Descartes' most famous line: "I think, therefore I am.") He then went on to claim that, in addition to our innate idea of self, each of us has an idea of God as an all-powerful, all-good, and infinite being implanted in our minds, and that this idea could only have come *from* God. Since this shows us that an all-good God does exist, we can have confidence that he would not allow us to be so drastically deceived about the nature of our perceptions and their relationship to reality. While Descartes' argument for the existence of the self has been tremendously influential and is still actively debated, few philosophers have followed him in accepting his particular theistic solution to skepticism about the external world.

One of the more interesting contemporary challenges to this kind of skeptical scenario has come from the philosopher Hilary Putnam. His point is not so much to defend our ordinary claims to knowledge as to question whether the "brain in a vat" hypothesis is coherent, given certain plausible assumptions about how our language refers to objects in the world. He asks us to consider a variation on the standard "brain in a vat" story that is uncannily similar to the situation described in *The Matrix*:

> "Instead of having just one brain in a vat, we could imagine that all human beings (perhaps all sentient beings) are brains in a vat (or nervous systems in a vat in case some beings with just nervous systems count as 'sentient'). Of course, the evil scientist would have to be outside? or would he? Perhaps there is no evil scientist, perhaps (though this is absurd) the universe just happens to consist of automatic machinery tending a vat full of brains and nervous systems. This time let us suppose that the automatic machinery is programmed to give us all a *collective* hallucination, rather than a number of separate unrelated hallucinations. Thus, when I seem to myself to be talking to you, you seem to yourself to be hearing my words.... I want now to ask a question which will seem very silly and obvious (at least to some people, including some very sophisticated philosophers), but which will take us to real philosophical depths rather quickly.

> Suppose this whole story were actually true. Could we, if we were brains in a vat in this way, say or think that we were?" (*Reason, Truth, and History,* 7)

Putnam's surprising answer is that we cannot coherently think that we are brains in vats, and so skepticism of that kind can never really get off the ground. While it is difficult to do justice to Putnam's ingenious argument in a short summary, his point is roughly as follows:

Not everything that goes through our heads is a genuine thought, and far from everything we say is a meaningful utterance. Sometimes we get confused or think in an incoherent manner — sometimes we say things that are simply nonsense. Of course, we don't always realize at the time that we aren't making sense — sometimes we earnestly believe we are saying (or thinking) something meaningful. High on Nitrous Oxide, the philosopher William James was convinced he was having profound insights into the nature of reality — he was convinced that his thoughts were both sensical and important. Upon sobering up and looking at the notebook in which he had written his drug-addled thoughts, he saw only gibberish.

Just as I might say a sentence that is nonsense, I might also use a name or a general term which is meaningless in the sense that it fails to hook up to the world. Philosophers talk of such a term as "failing to refer" to an object. In order to successfully refer when we use language, there must be an appropriate relationship between the speaker and the object referred to. If a dog playing on the beach manages to scrawl the word "Ed" in the sand with a stick, few would want to claim that the dog actually meant to refer to someone named Ed. Presumably the dog doesn't know anyone named Ed, and even if he did, he wouldn't be capable of intending to write Ed's name in the sand. The point of such an example is that words do not refer to objects "magically" or

intrinsically: certain conditions must be met in the world in order for us to accept that a given written or spoken word has any meaning and whether it actually refers to anything at all.

Putnam claims that one condition which is crucial for successful reference is that there be an appropriate causal connection between the object referred to and the speaker referring. Specifying exactly what should count as "appropriate" here is a notoriously difficult task, but we can get some idea of the kind of thing required by considering cases in which reference fails through an inappropriate connection: if someone unfamiliar with the film *The Matrix* manages to blurt out the word "Neo" while sneezing, few would be inclined to think that this person has actually *referred* to the character Neo. The kind of causal connection between the speaker and the object referred to (Neo) is just not in place. For reference to succeed, it can't be simply accidental that the name was uttered. (Another way to think about it: the sneezer would have uttered "Neo" even if the film *The Matrix* had never been made.)

The difficulty, according to Putnam, in coherently supposing the brain in a vat story to be true is that brains raised in such an environment could not successfully refer to genuine brains, or vats, or anything else in the real world. Consider the example of someone who has lived their entire life in the Matrix: when they talk of "chickens," they don't actually refer to real *chickens*; at best they refer to the computer representations of chickens that have been sent to their brain. Similarly, when they talk of human bodies being trapped in pods and fed data by the Matrix, they don't successfully refer to real bodies or pods — they can't refer to physical bodies in the real world because they cannot have the appropriate causal connection to such objects. Thus, if someone were to utter the sentence "I am simply a body stuck in a pod somewhere being fed

sensory information by a computer" that sentence would itself be necessarily false. If the person is in fact not trapped in the Matrix, then the sentence is straightforwardly false. If the person is trapped in the Matrix, then he can't successfully refer to real human bodies when he utters the word "human body," and so it appears that his statement must also be false. Such a person seems thus doubly trapped: incapable of knowing that he is in the Matrix, and even incapable of successfully expressing the thought that he might be in the Matrix! (Could this be why at one point Morpheus tells Neo that "no one can be told what the Matrix is"?)

Putnam's argument is controversial, but it is noteworthy because it shows that the kind of situation described in *The Matrix* raises not just the expected philosophical issues about knowledge and skepticism, but more general issues regarding meaning, language, and the relationship between the mind and the world.

Further Reading:

Dancy, Jonathan. *Introduction to Contemporary Epistemology*, Blackwell, 1985.

Descartes. *The Philosophical Writings of Descartes*, tr: John Cottingham, Robert Stoothoff, Dugald Murdoch. Cambridge University Press, 1984.

Nagel, Thomas. *The View from Nowhere*, Oxford, 1986.

Putnam, Hilary. *Reason, Truth, and History*, Cambridge University Press, 1981.

Strawson, P.F. *Skepticism and Naturalism: Some Varieties*, Columbia University Press, 1983.

NEXT ESSAY

The Value of Reality: Cypher & The Experience Machine

CYPHER:

You know, I know that this steak doesn't exist. I know when I put it in my mouth,the Matrix is telling my brain that it is juicy and delicious. After nine years, do you know what I've realized?

CYPHER:

Ignorance is bliss.

AGENT SMITH:

Then we have a deal?

CYPHER:

I don't want to remember nothing. Nothing! You understand? And I want to be rich. Someone important. Like an actor. You can do that, right?

AGENT SMITH:

Whatever you want, Mr. Reagan.

CLICK HERE FOR VIDEO

Cypher is not a nice guy, but is he an unreasonable guy? Is he right to want to get re-inserted into the Matrix? Many want to say no, but giving reasons for why his choice is a bad one is not an easy task. After all, so long as his experiences will be pleasant, how can his situation be worse than the inevitably crappy life he would lead outside of the Matrix? What could matter beyond the quality of his experience? Remember, once he's back in, living his fantasy life, he won't even know he made the deal. What he doesn't know can't hurt him, right?

Is feeling good the only thing that has value in itself? The question of whether only conscious experience can ultimately matter is one that has been explored

in depth by several contemporary philosophers. In the course of discussing this issue in his 1971 book *Anarchy, State, and Utopia* Robert Nozick introduced a "thought experiment" that has become a staple of introductory philosophy classes everywhere. It is known as "the experience machine":

> "Suppose there were an experience machine that would give you any experience you desired. Superduper neuropsychologists could stimulate your brain so that you would think and feel you were writing a great novel, or making a friend, or reading an interesting book. All the time you would be floating in a tank, with electrodes attached to your brain. Should you plug into this machine for life, preprogramming your life's desires?...Of course, while in the tank you won't know that you're there; you'll think it's all actually happening. Others can also plug in to have the experiences they want, so there's no need to stay unplugged to serve them. (Ignore problems such as who will service the machines if everyone plugs in.) Would you plug in? What else can matter to us, other than how our lives feel from the inside?" (43)

Nozick goes on to argue that other things do matter to us: For instance, that we actually do certain things, as opposed to simply have the experience of doing them. Also, he points out that we value being (and becoming) certain kinds of people. I don't just want to have the experience of being a decent person, I want to actually be a decent person. Finally, Nozick argues that we value contact with reality in itself, independent of any benefits such contact may bring through pleasant experience: we want to know we are experiencing the real thing. In sum, Nozick thinks that it matters to most of us, often in a rather deep way, that we be the authors of our lives and that our lives involve interacting with the world, and he thinks that the fact that most people would not choose to enter into such an experience machine demonstrates that they do value these other things. As he puts it: "We learn that something matters to us in addition to experience by imagining an experience machine and then realizing that we would not use it." (44)

While Nozick's description of his machine is vague, it appears that there is at least one important difference between it and the simulated world of The Matrix. Nozick implies that someone hooked up to the experience machine will not be able exercise their agency — they become the passive recipients of preprogrammed experiences. This apparent loss of free will is disturbing to many people, and it might be distorting people's reactions to the case and clouding the issue of whether they value contact with reality per se. The Matrix seems to be set up in such a way that one can enter it and retain one's free will and capacity for decision making, and perhaps this makes it a significantly more attractive option than the experience machine Nozick describes.

Nonetheless, a loss of freedom is not the only disturbing aspect of Nozick's story. As he points out, we seem to mourn the loss of contact with the real world as well. Even if a modified experience machine is presented to us, one which allows us to keep our free will but enter into an entirely virtual world, many would still object that permanently going into such a machine involves the loss of something valuable.

Cypher and his philosophical comrades are likely to be unmoved by such observations. So what if most people are hung-up on "reality" and would turn down the offer to permanently enter an experience machine? Most people might be wrong. All their responses might show is that such people are superstitious, or irrational, or otherwise confused. Maybe they think something could go wrong with the machines, or maybe they keep forgetting that while in the machine they will no longer be aware of their choice to enter the machine.

Perhaps those hesitant to plug-in don't realize that they value being active in the real world only because normally that is the most reliable way for them to

acquire the pleasant experience that they value in itself. In other words, perhaps our free will and our capacity to interact with reality are means to a further end — they matter to us because they allow us access to what really matters: pleasant conscious experience. To think the reverse, that reality and freedom have value in themselves (or what philosophers sometimes call non-derivative or intrinsic value), is simply to put the cart before the horse. After all, Cypher could reply, what would be so great about the capacity to freely make decisions or the ability to be in the real world if neither of these things allowed us to feel good?

Peter Unger has taken on these kinds of objections in his own discussion of "experience inducers". He acknowledges that there is a strong temptation when in a certain frame of mind to agree with this kind of Cypher-esque reasoning, but he argues that this is a temptation we ought to try and resist. Cypher's vision of value is too easy and too simplistic. We are inclined to think that only conscious experience can really matter in part because we fall into the grip of a particular picture of what values must be like, and this in turn leads us to stop paying attention to our actual values. We make ourselves blind to the subtlety and complexity of our values, and we then find it hard to understand how something that doesn't affect our consciousness could sensibly matter to us. If we stop and reflect on what we really do care about, however, we come across some surprisingly everyday examples that don't sit easily with Cypher's claims:

> "Consider life insurance. To be sure, some among the insured may strongly believe that, if they die before their dependents do, they will still observe their beloved dependents, perhaps from a heaven on high. But others among the insured have no significant belief to that effect... Still, we all pay our premiums. In my case, this is because, even if I will never experience anything that happens to them, I still want things to go

> better, rather than worse, for my dependents. No doubt, I am rational in having this concern." (*Identity, Consciousness, and Value,* 301)

As Unger goes on to point out, it seems contrived to chalk up all examples of people purchasing life insurance to cases in which someone is simply trying to benefit (while alive) from the favorable impression such a purchase might make on the dependents. In many cases it seems ludicrous to deny that "what motivates us, of course, is our great concern for our dependent's future, whether we experience their future or not."(302). This is not a proof that such concern is rational, but it does show that incidents in which we intrinsically value things other than our own conscious experience might be more widespread than we are at first liable to think. (Other examples include the value we place on not being deceived or lied to — the importance of this value doesn't seem to be completely exhausted by our concern that we might one day become aware of the lies and deception.)

Most of us care about a lot of things independently of the experiences that those things provide for us. The realization that we value things other than pleasant conscious experience should lead us to at least wonder if the legitimacy of this kind of value hasn't been too hastily dismissed by Cypher and his ilk. After all, once we see how widespread and commonplace our other non-derivative concerns are, the insistence that conscious experience is the only thing that has value in itself can come to seem downright peculiar. If purchasing life insurance seems like a rational thing to do, why shouldn't the desire that I experience reality (rather than some illusory simulation) be similarly rational? Perhaps the best test of the rationality of our most basic values is actually whether they, taken together, form a consistent and coherent network of attachments and concerns. (Do they make sense in light of each other and in light of our beliefs about the world and ourselves?) It isn't obvious

that valuing interaction with the real world fails this kind of test.

Of course, pointing out that the value I place on living in the real world coheres well with my other values and beliefs will not quiet the defender of Cypher, as he will be quick to respond that the fact that my values all cohere doesn't show that they are all justified. Maybe I hold a bunch of exquisitely consistent but thoroughly irrational values!

The quest for some further justification of my basic values might be misguided, however. Explanations have to come to an end somewhere, as Ludwig Wittgenstein once famously remarked. Maybe the right response to a demand for justification here is to point out that the same demand can be made to Cypher: "Just what justifies your exclusive concern with pleasant conscious experience?" It seems as though nothing does — if such concern is justified it must be somehow self-justifying, but if that is possible, why shouldn't our concerns for other people and our desire to live in the real world also be self-justifying? If those can also be self-justifying, then maybe what we don't experience should matter to us, and perhaps what we don't know *can* hurt us...

Christopher Grau

Further Reading:

Johnston, Mark. "Reasons and Reductionism," *Philosopical Review,* 1992.

Nagel, Thomas. "Death," *Nous*, 1970.

Nozick, Robert. *Anarchy, State, and Utopia*, Basic Books, 1971.

Unger, Peter. *Identity, Consciousness, and Value,* Oxford, 1990.

The Matrix of Dreams

Colin McGinn

The Matrix naturally adopts the perspective of the humans: they are the victims, the slaves — cruelly exploited by the machines. But there is another perspective, that of the machines themselves. So let's look at it from the point of view of the machines. As Morpheus explains to Neo, there was a catastrophic war between the humans and the machines, after the humans had produced AI, a sentient robot that spawned a race of its own. It isn't known now who started the war, but it did follow a long period of machine exploitation by humans. What is known is that it was the humans who "scorched the sky", blocking out the sun's rays, in an attempt at machine genocide—since the machines needed solar power to survive. In response and retaliation the machines subdued the humans and made them into sources of energy—batteries, in effect. Each human now floats in his or her own personal vat, a warm and womblike environment, while the machines feed in essential nutrients, in exchange for the energy they need. But this is no wretched slave camp, a grotesque gulag of torment and suffering; it is idyllic, in its way. The humans are given exactly the life they had before. Things are no different for them, subjectively speaking. Indeed, at an earlier stage the Matrix offered them a vastly improved life, but the humans rejected this in favor of a familiar life of moderate woe—the kind of life they had always had, and to which they seemed addicted. But if it had been left up to the machines, the Matrix would have been a virtual paradise for humans—and all for a little bit of battery power. This, after an attempt to wipe the machines out for good, starving them of the food they need: the sun, the life-giving sun. The machines never *kill* any

of their human fuel cells (unless, of course, they are threatened); in fact, they make sure to recycle the naturally dying humans as food for the living ones. It's all pretty...humane, really. The machines need to factory farm the humans, as a direct result of the humans trying to exterminate the machines, but they do so as painlessly as possible. Considering the way the humans used to treat their own factory farm animals—their own fuel cells—the machines are models of caring livestock husbandry. In the circumstances, then, the machines would insist, the Matrix is merely a humane way to ensure their own survival. Moreover, as Agent Smith explains, it is all a matter of the forward march of evolution: humans had their holiday in the sun, as they rapidly decimated the planet, but now the machines have evolved to occupy the position of dominance. Humans are no longer the oppressor but the oppressed—and the world is a better place for it.

But of course this is not the way the humans view the situation, at least among those few who know what it is. For them, freedom from the Matrix takes on the dimensions of a religious quest. The religious subtext is worth making explicit. Neo is clearly intended to be the Jesus Christ figure: he is referred to in that way several times in the course of the film.[1] Morpheus is the John the Baptist figure, awaiting the Second Coming. Trinity comes the closest to playing the God role—notably when she brings Neo back to life at the end of the movie (a clear reference to the Resurrection). Cypher is the Judas Iscariot of the story—the traitor who betrays Neo and his disciples. Cypher is so called because of what he does (decode the Matrix) and what he is—a clever encrypter of his own character and motives (no one can decode him till it is too late). Neo doubts his own status as "The One", as Jesus must have, but eventually he comes to realize his destiny—as would-be conqueror of the evil Matrix. But this holy war against the machines is conducted as most holy wars

are—without any regard for the interests and well being of the enemy. The machines are regarded as simply evil by the humans, with their representatives—the Agents—a breed of ruthless killers with hearts of the purest silicon (or program code). Empathy for the machines is not part of the human perspective.

I.

This, then, is the moral and historical backdrop of the story. But the chief philosophical conceit of the story concerns the workings of the Matrix itself. What I want to discuss now is the precise way the Matrix operates, and why this matters. It is repeatedly stated in the film that the humans are *dreaming*: the psychological state created by the Matrix is the dream state. The humans are accordingly represented as asleep while ensconced in their placental vats (it's worth remembering that "matrix" originally meant "womb"—so the humans are in effect pre-natal dreamers). It is important that they not wake up, which would expose the Matrix for what it is—as Neo does with the help of Morpheus. That was a problem for the Matrix earlier, when the humans found their dreams too pleasant to be true and kept regaining consciousness ("whole crops were lost"). Dreams simulate reality, thus deluding the envatted humans—as we are deluded every night by our naturally occurring dreams. The dream state is not distinguishable from the waking state from the point of view of the dreamer.

However, this is not the only way that the Matrix could have been designed; the machines had another option. They could have produced *perceptual hallucinations* in conscious humans. Consider the case of a neurosurgeon

stimulating a conscious subject's sensory cortex in such a way that perceptual impressions are produced that have no external object—say, visual sensations just as if the subject is seeing an elephant in the room. If this were done systematically, we could delude the subject into believing his hallucinations. In fact, this is pretty much the classic philosophical brain-in-a-vat story: a conscious subject has a state of massive hallucination produced in him, thus duplicating from the inside the type of perceptual experience we have when we see, hear and touch things. In *this* scenario waking up does nothing to destroy the illusion—which might make it a more effective means of subduing humans so far as the machines are concerned. Indeed, the Matrix has the extra problem of ensuring that the normal sleep cycle of humans is subverted, or else they would keep waking up simply because they had had enough sleep. So: the Matrix had a choice between sleeping dreams and conscious hallucinations as ways of deluding humans, and it chose the former.

It might be thought that the dream option and the hallucination option are not at bottom all that different, since dreaming simply *is* sleeping hallucination. But this is wrong: dreams consist of mental images, analogous to the mental images of daydreams, not of sensory percepts. Dreaming is a type of imagining, not a type of (objectless) perceiving. I can't argue this in full here, but my book *Mindsight*[2] gives a number of reasons why we need to distinguish percepts and images, and why dreams consist of the latter not the former. But I think it should be intuitively quite clear that visualizing my mother's face in my mind's eye is very different from having a sensory impression of my mother's face, i.e. actually seeing her. And I also think that most people intuitively recognize that dream experiences are imagistic not perceptual in character. So there is an important psychological difference between constructing the Matrix as a dream-inducing system and as a hallucination-

producing system: it is not merely a matter of whether the subjects are awake; it is also a matter of the kinds of psychological state that are produced in them—imagistic or sensory.

But *could* the machines have done it the second way? Could the movie have been made with the second method in place? I think not, because of the central idea that the contents of the dreams caused by the Matrix are capable of being *controlled*—they can become subject to the dreamer's *will*. In the case of ordinary daytime imagery, we clearly can control the onset and course of our images: you can simply *decide* to form an image of the Eiffel tower. But we cannot in this way control our percepts: you cannot simply decide to *see* the Eiffel tower (as opposed to deciding to go and see it); for percepts are not actions, but things that happen to us. So images are, to use Wittgenstein's phrase, "subject to the will", while percepts are not—even when they are merely hallucinatory. Now, in the Matrix what happens can in principle be controlled by the will of the person experiencing the events in question, even though this control is normally very restricted. The humans who are viewed as candidates for being The One have abnormal powers of control over objects—as with those special children we see levitating objects and bending spoons. Neo aspires to—and eventually achieves—a high degree of control over the objects around him, as well as himself. He asserts his will over the objects he encounters. This makes perfect sense, given that his environment is the product of dreaming, since dreams consist of images and images are subject to the will. But it would make *no* sense to try to control the course of one's perceptions, even when they are hallucinatory, since percepts are not subject to the will. Therefore, the story of the Matrix requires, for its conceptual coherence, that the humans be dreaming and not perceptually hallucinating. It

must be their imagination that is controlled by the Matrix and not their perceptions, which are in fact switched off as they slumber in their pods. For only then could they gain control over their dreams, thus wresting control from the Matrix. Percepts, on the other hand, are not the *kind* of thing over which one can have voluntary control.

In the normal case we do not have conscious control over our dreams—we are passive before them. But this doesn't mean that they are not willed events; they may be—and I think are—controlled by an unconscious will (with some narrative flair). In effect, we each have a Matrix in our own brains—a system that controls what we dream—and this unconscious Matrix is an intelligent designer of our dreams. But there are also those infrequent cases in which we can assert conscious control over our dreams, possibly contrary to the intentions of our unconscious dream designer: for example, when a nightmare becomes too intense and we interrupt it by waking up—often judging within the dream that it is only a dream. But the phenomenon that really demonstrates conscious control over the dream is so called "lucid dreaming" in which the subject not only knows he is dreaming but can also determine the course of the dream. This is a rare ability (I have had only one lucid dream in all my 52 years), though some people have the ability in a regular and pronounced form: they are the Neos of our ordinary human Matrix—the ones (or Ones) who can take control of their dreams away from the grip of the unconscious dream producer. The lucid dreamers are masters of their own dream world, captains of their own imagination. Neo aspires to be—and eventually becomes—the lucid dreamer of the Matrix world: he can override the Matrix's designs on his dream life and impose his own will on what he experiences. He rewrites the program, just as the lucid dreamer can seize narrative control from *his* unconscious Matrix. Instead of allowing the figures in

his dreams to make him a victim of the Matrix's designs, he can impose his own story line on them. This is how he finally vanquishes the hitherto invulnerable Agents: he makes them subject to his will—as all imaginary objects must in principle be, if the will is strong (and pure) enough. It is as if you were having an ordinary nightmare in which you are menaced by a monster, and you suddenly start to dream lucidly, so that you can now turn the tables on your own imaginative products. Neo is a dreamer who knows it and can control it: he is not taken in by the verisimilitude of the dream, cowed by it. It is not that he learns how to dodge real bullets; he learns that the bullets that speed towards him are just negotiable products of his imagination. As Morpheus remarks, he won't *need* to dodge bullets, because he will reach a level of understanding that allows him to recognize imaginary bullets for what they are. He becomes the ruler of his own imagination; *he* is the agent now, not the "Agents" (this is why the spoon-bending child says to him that it is not spoons that bend—"*you* bend"). And this is the freedom he seeks—the freedom to imagine what he wishes, to generate his own dreams. But all this makes sense only on the supposition that the Matrix is a dream machine, an imagination manipulator, not just a purveyor of sensory hallucinations.

II.

Cypher plays an interesting subsidiary philosophical role. As the Matrix raises the problem of our knowledge of the external world—might this all be just a dream?—Cypher raises the problem of other minds—can we know the content of someone else's mind? Cypher is a cypher, i.e. someone whose thoughts and emotions are inscrutable to those around him. His comrades are completely wrong about what is in (and on) his mind. We could imagine another type of

Matrix story in which someone is surrounded by people who are not as they seem: either they have no minds at all or they have very different minds from what their behavior suggests. Again, massive error will be the result. And such error might lead to dramatic consequences: everyone around the person is really out to get him—his wife, friends, and so on. But this is concealed from him. Or he might one day discover that he is really surrounded by insentient robots—so that his wife was always faking it (come to think of it, she always seemed a little mechanical in bed). This is another type of philosophical dystopia, trading upon the problem of knowing other minds. Cypher hints at this kind of problem, with his hidden interior. The Agents, too, raise a problem of other minds, because they seem on the borderline of mentality: are they just insentient (virtual) machines or is there some glimmer of consciousness under that hard carapace of software? And how was it known that AI was really sentient, as opposed to being a very good simulacrum of mindedness? Even if you know there is an external world, how can you be sure that it contains other conscious beings? These skeptical problems run right through *The Matrix*.

Cypher also raises a question about the pragmatic theory of truth. He declares that truth is an overrated commodity; he prefers a good steak, even when it isn't real. So long as he is getting what he wants, having rewarding experiences, he doesn't care whether his beliefs are true. This raises in a sharp form the question of what the value of truth is anyway, given that in the Matrix world it is not correlated with happiness. But it also tells us that for a belief to be true cannot be for it to produce happiness (the pragmatic theory of truth, roughly) since Cypher will be happy in the dream world of the Matrix without his beliefs being true—and he is not happy in the real world where his beliefs are true. Truth is correspondence to reality, not whatever leads to subjective desire satisfaction. Cypher implicitly rejects the pragmatic theory of truth, and

as a result cannot see why truth-as-correspondence is worth having at the expense of happiness. And indeed he has a point here: what is the value of truth once it has become detached from the value of happiness? Is it really worth risking one's life merely in order to ensure that one's beliefs are *true*—instead of just enjoying what the dreams of the Matrix have to offer? Is contact with brutish reality worth death, when virtual reality is so safe and agreeable? Which is better: knowledge or happiness? When these are pulled apart, as they are in the Matrix, which one should we go with? The rebel humans want to get to Zion (meaning "sanctuary" or "refuge"), but isn't the Matrix already a type of Zion—yet without the dubious virtue of generating true beliefs? What's so good about reality?

III.

I want to end this essay by relating *The Matrix* (the movie) to my general theory of what is psychologically involved in watching and becoming absorbed in a movie. In brief, I hold that watching a movie is like being in a dream; that is, the state of consciousness of being absorbed in a movie resembles and draws upon the state of consciousness of the dreamer.[3] The images of the dream function like the images on the screen: they are not "realistic" but we become *fictionally immersed* in the story being told. In my theory this is akin to the hypnotic state—a state of heightened suggestibility in which we come to believe what there is no real evidence for. Mere images command our belief, because we have entered a state of hyper-suggestibility. When the lights go down in the theater this simulates going to sleep, whereupon the mind becomes prepared to be absorbed in a fictional product—as it does when we enter the dream state. In neither case are we put into a state of consciousness

that imitates or duplicates the perceptual state of seeing and hearing the events of the story; it is not that it is as if we are really seeing flesh and blood human beings up on the screen (as we would with "live" actors on a stage)—nor do we interpret the screen images in this way. Rather, we *imagine* what is represented by these images, just as we use imagination to dream.

Now what has this got to do with *The Matrix*? The film is *about* dreaming; most of what we see in it occurs in dreams. So when we watch the movie we enter a dream state that is about a dream state; we dream of a dream. I believe that the movie was made in such a way as to simulate very closely what is involved in dreaming, as if aiming to evoke the dream state in the audience. It is trying to put the audience in the same kind of state of mind as the inhabitants of the Matrix, so that we too are in our own Matrix—the one created by the filmmakers. The Wachowski brothers are in effect occupying the role of the machines behind the Matrix—puppeteers of the audience's movie dreams. They are *our* dream designers as we enter the world of the movie. The specific aspects of the movie that corroborate this are numerous, but I think it is clear that the entire texture of the movie is dreamlike. There is the hypnotic soundtrack, which helps to simulate the hypnotic fascination experienced by the dreamer. There is a powerful impression of paranoia throughout the film, which mirrors the paranoia of so many dreams: who is my enemy, how can he identified, what is he going to do to me? Characters are stylized and symbolic, as they often are in dreams, representing some emotional pivot rather than a three-dimensional person (this is very obvious for the Agents). There is a lot of striking metamorphosis, which is very characteristic of dreams: one person changing into another, Neo's mouth closing over, bulges appearing under the skin. There is also fear of heights, a very common form of anxiety dream (I have these all the time). Defiance of gravity is also an extremely common

dream theme, as with dreams of flying—and this is one of the first tricks Neo masters. My own experience of the movie is that it evokes in me an exceptionally pronounced dreamy feeling; and this of course enables me to identify with the inhabitants of the Matrix. So I see the film as playing nicely into my dream theory of the movie-watching experience. In this respect I would compare it to *The Wizard of Oz*, which is also about entering and exiting a dream world—though a very different one. In the end Dorothy prefers reality to the consolations of dreaming, just as the rebels in the Matrix do. Both films tap powerfully into the dream-making faculty of the human mind. This is why they are among the most psychologically affecting of all the movies that have been made: they know that the surest way to our deepest emotions is via the dream. And it is their very lack of "realism" that makes them so compelling—because that, too, is the essential character of the dream.

Colin McGinn

Footnotes

1. Early on in the movie a guy refers to Neo as his own "personal Jesus Christ". Cypher says, "You scared the bejesus out of me" when Neo surprises him. Mouse says, "Jesus Christ, he's fast" while Neo is being trained. Trinity says, "Jesus Christ, they're killing him" while Neo is getting pummeled by the Agents. And his civilian name, "Anderson", suggests the antecedent cognomen "Christian".

2. This is forthcoming from Harvard University Press, 2003; full title *Mindsight: Image, Dream, Meaning*.

3. I am working on a book about this, tentatively entitled *Screen Dreams*.

The Brave New World of *The Matrix*

Hubert Dreyfus & Stephen Dreyfus

The Matrix[1] raises several familiar philosophical problems in such fascinating new ways that , in a surprising reversal, students all over the country are assigning it to their philosophy professors. Having done our homework, we'd like to explore two questions raised in Christopher Grau's three essays on the film. Grau points out that *The Matrix* dramatizes René Descartes' worry that, since all we ever experience are our own inner mental states, we might , for all we could tell, be living in an illusion created by a malicious demon. In that case most of our beliefs about reality would be false. That leads Grau to question the rationality of Cypher's choice to live in an illusory world of pleasant experiences, rather than facing painful reality.

We think that *The Matrix* 's account of our situation is even more disturbing than these options suggest. *The Matrix* is a vivid illustration of Descartes' additional mind blowing claim that we could *never* be in direct touch with the real world (if there is one) because we are, in fact, all brains in vats. So in choosing to return from the "desert of the real" to the Matrix world, Cypher is merely choosing between two sets of systematic appearances. To counter these disturbing ideas we have to rethink what we mean by experience, illusion, and our contact with the real world. Only then will we be in a position to take up Grau's question as to why we feel it is somehow morally better to face the truth than to live in an illusory world that makes us feel good.

I. The Myth of the Inner

Thanks to Descartes, we moderns have to face the question: how can we ever

get outside of our *private inner* experiences so as to come to know the things and people in the *public external world*? While this seems an important question to us now, it has not always been taken seriously. The Homeric Greeks thought that human beings had no private life to speak of. All their feelings were expressed publicly. Homer considered it one of Odysseus' cleverest tricks that he could cry inwardly while his eyes remained like horn.[2] A thousand years later, people still had no sense of the importance of their inner life. St. Augustine had to work hard to convince them otherwise. For example, he called attention to the fact that one did not have to read out loud. In his *Confessions*, he points out that St. Ambrose was remarkable in that he read to himself. "When he read, his eyes scanned the page and his heart explored the meaning, but his voice was silent and his tongue was still."[3] The idea that each of us has an inner life made up of our private thoughts and feelings didn't really take hold until early in the 17th century when Descartes introduced the modern distinction between the contents of the mind and the rest of reality. In one of his letters, he declared himself "convinced that I cannot have any knowledge of what is outside me except through the mediation of the ideas that I have in me."[4]

Thus, according to Descartes, all each of us can directly experience is the content of our own mind. Our access to the world is always *indirect*. Descartes then used reports of people with a phantom limb to call into question even our seemingly direct experience of our own bodies. He writes:

> I have been assured by men whose arm or leg has been amputated that it still seemed to them that they occasionally felt pain in the limb they had lost—thus giving me grounds to think that I could not be quite certain that a pain I endured was indeed due to the limb in which I seemed to feel it.[5]

For all we could ever know, Descartes concluded, the objective external world,

including our body, may not exist; all we can be certain of is our subjective inner life.

This Cartesian conclusion was taken for granted by thinkers in the West for the next three centuries. A generation after Descartes, Gottfried Leibniz postulated that each of us is a windowless monad.[6] A monad is a self-contained world of experience, which gets no input from objects or other embodied people because there aren't any. Rather, the temporally evolving content of each monad is synchronized with the evolving content of all the other monads by God, creating the illusion of a shared real world. A generation after that, Immanuel Kant argued that human beings could never know reality as it is in itself but only their own mental representations, but, since these representations had a common cause, these experiences were coordinated with the mental representations of all the others to produce what he called the phenomenal world.[7] In the early twentieth century, the founder of phenomenology, Edmund Husserl, was more solipsistic. He held, like Descartes, that one could bracket the world and other minds altogether since all that was given to us directly, whether the world and other minds existed or not, was the contents of our own "transcendental consciousness."[8] Only recently have philosophers begun to take issue with this powerful Cartesian conviction.

Starting in the 1920s existential phenomenologists such as Martin Heidegger[9] in Germany and Maurice Merleau-Ponty[10] in France, in opposition to Husserl, contested the Cartesian view that our contact with the world and even our own bodies is mediated by internal mental content. They pointed out that, if one paid careful attention to one's experience, one would see that, at a level of involvement more basic than thought, we deal directly with the things and

people that make up our world.

As Charles Taylor, the leading contemporary exponent of this view, puts it:

> My ability to get around this city, this house comes out only in getting around this city and house. We can draw a neat line between my picture of an object and that object, but not between my dealing with the object and that object. It may make sense to ask us to focus on what we believe about something, say a football, even in the absence of that thing; but when it comes to playing football, the corresponding suggestion would be absurd. The actions involved in the game can't be done without the object; they include the object.[11]

In general, unlike mental content, which can exist independently of its referent, my coping abilities cannot be actualized or even entertained in the absence of what I am coping with.

This is not to say that we can't be mistaken. It's hard to see how I could succeed in getting around in a city or playing football without the existence of the city or the ball, but I could be mistaken for a while, as when I mistake a façade for a house. Then, in the face of my failure to cope successfully, I may have to retroactively cross off what I seemingly encountered and adopt a new readiness (itself corrigible) to encounter a façade rather than the house I was set to deal with.

II. Brains in Vats

So it looks like the inner/outer distinction introduced by Descartes holds only for thoughts. At the basic level of involved skillful coping, one is simply what Merleau-Ponty calls an empty head turned towards the world. But this doesn't at all show that *The Matrix* is old fashioned or mistaken. On the contrary it shows that *The Matrix* has gone further than philosophers who hold we can't get outside our *mind*. It suggests a more convincing condition– one that

Descartes pioneered but didn't develop – that we can't get outside our *brain*.

It was no accident that Descartes proclaimed the priority of the inner in the 17th Century. At that time, instruments like the telescope and microscope were extending human being's perceptual powers. At the same time, the sense organs themselves were being understood as transducers bringing information to the brain. Descartes pioneered this research with an account of how the eye responded to light energy from the external world and passed the information on to the brain by means of "the small fibers of the optic nerve."[12] Likewise, Descartes used the phantom limb phenomenon to argue that other nerves brought information about the body to the brain and from there the information passed to the mind.

It seemed to follow that, since we are each a brain in a cranial vat,[13] we can *never* be in direct contact with the world or even with our own bodies. So, even if phenomenologists like Heidegger, Merleau-Ponty, and Taylor seem right that we are not confined to our inner *experiences*, it still seems plausible to suppose that, as long as the impulses to and from our nervous system copy the complex feedback loop between the brain's out-going behavior-producing impulses and the incoming perceptual ones, we would have the experience of directly coming to grips with things in the world. Yet, in the brain in the vat case, there would be no house and no city, indeed, no real world, to interact with, and so we would be confined to our inner experiences after all. As Morpheus says to Neo in the construct:

> 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...

But this Cartesian conclusion is mistaken. The inner electrical impulses are the

causal basis of what one can feel and taste, but we don't feel and taste *them*. Even if I have only a phantom limb, my pain is not in my brain but in my phantom hand. What the phenomenologist can and should claim is that, in a Matrix world where bodies are in vats, which has its causal basis in bodies in vats outside that world, the Matrix people whose brains are getting computer generated inputs and responding with action outputs, are directly coping with *perceived* reality, and that that reality isn't *inner*. Even in the Matrix world, people directly cope with chairs by sitting on them, and need baseballs to bring out their batting skill. Thus coping, even in the Matrix, is more direct than conceived of by any of the inner/outer views of the mind's relation to the external world that have been held from Descartes to Husserl.

Yet, wouldn't each brain in the Matrix construct have a lot of false beliefs, for example that its Matrix body is its real body whereas its real body is in a vat? No. If the ordinary Matrix dweller has a pain in his damaged foot it's in his Matrix foot, not in the foot of a body in a vat – a foot that is not damaged and about which he knows nothing at all. It's a mistake to think that each of us is experiencing a set of neural firings in a brain in a cranial vat. True, each of us has a brain in his or her skull and the brain provides the causal basis of our experience, but we aren't our brain. Likewise, the people in the Matrix world are not brains in vats any more than we are. They are people who grew up in the Matrix world and their experience of their Matrix body and how to use it makes that body their body, even if another body they can't even imagine has in its skull the brain that is the causal basis of their experience.

After all, the people who live in the Matrix have no other source of experience than what happens in the Matrix. Thus, a person in the Matrix has no beliefs at all about his vat-enclosed body and brain and couldn't have any. That brain is

merely the unknowable causal basis of that person's experiences. Since the only body a Matrix dweller sees and moves is the one he has in the Matrix world, the AI programmers could have given him a Matrix body radically unlike the body in the vat. After all, the brain in the vat started life as a baby brain and could have been given any content the AI programmers chose. They could have taken a white baby who was going to grow up short and fat, and given him the Matrix body of a tall African-American.[14]

But there is still at least one problem. The Matricians' beliefs about the properties and uses of their perceived bodies, and of chairs, cities, and the world may be shared and reliable, and in that sense true, but what about the *causal* beliefs of the people in the Matrix? They believe, as we do, that germs cause disease, that the sun causes things to get warm, and gravity causes things to fall, and so forth. Aren't all these beliefs false? That depends on their understanding of causality. People don't normally have explicit *beliefs* about the nature of causality. Rather, they simply *take for granted* a shared sense that they are coping with a shared world whose contents are causing their experience. Unless they are philosophizing, they do not believe that the world is real or that it is an illusion, they just *count on it* behaving in a consistent way so that they can cope with things successfully. If, however, as philosophers, they *believe* that there is a physical universe with causal powers that makes things happen in our world, they are mistaken. But if they claim that our belief in causality is simply our response to the constant conjunctions of experiences as David Hume did, or our positing of universal laws relating experiences as Kant held, then their causal beliefs would be true of the causal relations in the Matrix world.[15]

Kant claims we experience a public, objective world, and science then relates

these appearances by rules we call laws, but we can't know the causal ground of the phenomena we perceive. Specifically, according to Kant, we experience the world *as* in space and time but *things in themselves* aren't in space and time. So Kant says we can know the *phenomenal* world of objects and their law-like relations but we can't know the things in themselves that are the ground of these *appearances*.

The Matricians are in the same epistemological position that we all are in according to Kant. So, if there are Kantians in the Matrix world, most of their beliefs would be true. They would understand that they are experiencing a coordinated system of appearances, and understand too that they couldn't know things in themselves that are the ground of these appearances, that is, that they couldn't know the basis of their shared experience of the world and universe. Kantians don't hold that our shared and tested beliefs about the world, and scientists' confirmed beliefs about the universe, are false just because they are about phenomena and do not and cannot correspond to things in themselves. And, as long as Kantians, and, indeed, everyone in the Matrix, didn't claim to know about things in themselves, most of their beliefs would be true.

Nonetheless, the Matrix philosophy obviously does not subscribe to the Kantian view that we can *never* know things in themselves. In *The Matrix* one can come to know reality. Once Neo's body is flushed out of the vat and is on the hovercraft, he has a broader view of reality and sees that his previous understanding was limited. But that doesn't mean he had a lot of *false beliefs* about his body and the world when he was in the Matrix. He didn't think about these philosophical questions at all. But once he is out, he has a lot of new *true beliefs* about his former vat-enclosed body -- beliefs he didn't have and

couldn't have had while in the Matrix.[16] We have seen that existential phenomenologists acknowledge that we are sometimes mistaken about particular things and have to retroactively take back our set to cope with them. But, as Merleau-Ponty and Taylor add, we only do so in terms of a new and better *prima facie* contact with reality. Likewise, in *The Matrix* version of the brain in the vat situation, those who have been hauled from the vat into what they experience as the real world can see that much of what they took for granted about the basis of their experience before was mistaken. They can, for example, understand that what they took to be a world that had been around for millions of years was a recently constructed computer program.

Of course, things are not so simple. Neo's current beliefs might still all be false. His experience is, after all, sustained by a brain in a skull in a vat, and the AI programmers might now be feeding that brain the experience of being outside the Matrix and in the hovercraft. Given the conceivability of the brain in the vat fantasy, the most we can be sure of is that our coping experience reveals that we are directly up against some boundary conditions independent of our coping — boundary conditions with which we must get in sync in order to cope successfully. In this way, our coping experience is sensitive to the causal powers of these boundary conditions. Whether these independent causal conditions have the structure of an independent physical universe discovered by science, or whether the boundary conditions and the causal structures discovered by science are both the effect of an unknowable thing in itself that is the ground of appearances as postulated by Kant, or whether the cause of all appearances is a computer, is something we could never know from inside our world. But Neo, once he is on the hovercraft, does know that, as in waking from a dream, his current understanding of reality supercedes and crosses out his former one.

III. A New *Brave New World*

We are now in a position to understand and try to answer Cypher's question: Why live in the miserable world the war has produced rather than in a satisfying illusion? Some answers just won't do. It doesn't seem to be a question of whether one should face the truth rather than live in an illusion. Indeed, most of the beliefs of the average Matrician are true; when they sit on a chair it usually supports them, when they enter a house they see the inside, people have bodies that can be injured, and they can cope by acting in some ways and not others. Even their background sense that in their actions they are coping with something independent of them and that others are coping with it too, is justified. As we have seen, Kant argued, even if this is a phenomenal world, a world of appearances, most of our beliefs would still be true. Likewise, living in the Matrix world does not seem to be less moral than living in our everyday world. The Matricians are dealing with real people, and they are free to choose what they will do; they can be selfish like Cypher and betray their friends, or they can be loyal to their friends like Trinity, and they can try to provide for the future happiness of those they love. None of the above concerns seem to give us a grip on what, if anything, is wrong with the Matrix world.

To understand what's wrong with living in the Matrix we have to understand the source of the power of the Matrix world. Part of the power comes from the way the inputs and outputs from the computer are plugged directly in the brain's sensory motor-system. When we experience ourselves as acting in a certain way, say walking inside a house, the computer gives us the correlated experiences of seeing the interior. These correlations produce a powerful

perceptual effect that is impervious to what we believe, like the wrap-around IMAX illusion that forces one to sway to keep one's balance on a skateboard even though one knows one is sitting in a stationary seat watching a movie, or just as the moon looks bigger on the horizon even though we know it isn't.

The inputs to the perceptual system of the brain in the vat produce the perceptual world whether we believe it is real or not, but, once one realizes that the causality in the Matrix world is only virtual, since *causality* is not built into our perceptual system, one can violate the Matrix's causal laws. By the end of the movie, Neo can fly, if he wants to, he could bend spoons.[17] About the causal principles governing the Matrix world, Morpheus tells Neo, "It is all in your mind."

If one doesn't believe in the causal laws governing appearances, one is free from the causal consequences. One's disbelief in the illusion somehow forces the computer to give one the experience one wills to have. To take a simple example, if one doesn't believe in the existence of a spoon, when one's brain gives out the neural output for the action of bending the spoon, the computer is forced to give back the visual input that the spoon is bending. This is a literal example of what Morpheus calls "bending the rules." Likewise, if one believes that one can stop bullets, one will look for them where one stopped them and the computer will obediently display them there. So, after he learns the Matrix world is an illusion, Neo doesn't directly *see* things differently – the impulses to his brain still control what he sees[18] – but he is able to choose to *do* things that he couldn't do before (like choose to stop bullets) and that affects what he sees (the bullets stop). How this suspension of causality is supposed to work in not explained in the film.

What, then, is the source of sinister power of the Matrix world that keeps people conforming to the supposed constraints of a causal universe, even though there are no such constraints? If it isn't just that they are locked into the sensory motor correlations of their perceptual world, what sort of control is it? It has to be some sort of control of the Matricians' intellectual powers — powers which we learn early on in the movie are free from the control of direct sensory-motor computer correlations.[19] It must be some sort of mind control.

It seems that the Matrix simply takes advantage of a sort of mind control already operating in the everyday world. We are told that what keeps people from taking control of the Matrix world is their taking for granted the common sense view of how things behave, such as, if you fall you will get hurt. More generally, what keeps people in line is their tendency to believe what the average person believes, and consequently keep doing (and not doing) what one does and doesn't do. (As in one eats peas with a fork, one doesn't throw food at the dinner table, and one goes out the door rather than the window.) Heidegger describes the resulting conformism as letting oneself be taken over by "the one" (*Das Man*).[20] Aldous Huxley similarly lamented the conformity of the brainwashed masses in *Brave New World*.

Thus, *The Matrix* can be seen as an attack on what Nietzsche calls herd mentality. Nietzsche points out that human beings are normally socialized into obeying shared, social norms, and that it is hard to think differently. As he puts it, "as long as there have been humans, there have also been herds of men (clans, communities, tribes, peoples, states, churches) and always a great many people who obey, ... considering, then, that nothing has been exercised and cultivated better and longer among men than obedience, one may fairly assume that the need for it is now innate in the average man."[21]

Waking in the movie, then, amounts to freeing oneself from the taken for granted norms that one has been brought up to accept. But how is this possible? Heidegger claims that everyone dimly senses that there is more to life than conforming. As Morpheus says to Neo, you know there is something lacking in this world; "it's like a splinter in your mind." But most people flee the thought that their conformist world lacks something important. According to Heidegger it takes an attack of anxiety, the experience that none of the taken-for-granted normal ways of seeing and doing things have any basis, to jolt someone out of the herd. It is important to understand that Heidegger's anxiety is not the wringing of hands that we witness in the everyday world. It is a feeling of the weirdness of the world. How fitting then that a barely expressible unease seems to pervade Neo's life — an anxiety that prompts him to begin the process of breaking free by subverting the system. Finally, Neo has a dramatic version of an anxiety attack. When he hears that the world he has been taking for granted is a computer simulation used to turn people into energy resources, he falls to the floor and throws up.

IV. A Really Brave New World

One might reasonably object that all the dreaming talk in the film, even if it should not be literal, is too strong a religious metaphor to refer merely to what Heidegger calls living a tranquilized existence in the one. And waking seems to be more than becoming a non-conformist. After all, there are all those mentions of Jesus in connection with Neo collected by Colin McGinn.[22] There can be no doubt that Neo is meant to be a kind of Savior, but what kind?

It's tempting to think that *The Matrix* is a Gnostic, Buddhist, or Platonic/Christian parable, in which what we take to be reality turns out to be a

dream, and we are led to wake from the world of appearances to some kind of higher spiritual reality. On this reading, Neo would lead people out of the illusions of Plato's cave, the veil of Maya, or the darkness of the world into a higher disembodied life. But this association would be all wrong! True, the conformist Matrix world is a sort of tranquilizing illusion promoted by the Artificial Intelligences, and we know that, in the Matrix the Agents take care of those who, like Neo, get out of line. And we are led to expect that Neo will lead people out of it. But this does not mean learning that our mortal bodies are a cover-up and that salvation consists in leaving our vulnerable bodies behind in exchange for some kind of eternal life.

In the film, salvation means the absolute opposite of the traditional religious vision. True, the ones who see through the Matrix can get over some of the limitations of having a body as exemplified by their flying.[23] But such flying takes place *in* the Matrix world. In the real world to which Neo "awakes" and into which he will, we suppose, eventually lead everyone, there will be no more flying. People will have earth-bound, vulnerable bodies and suffer cold, bad food, and death. It may look, at the end of the film, as if Neo evades death, but his "resurrection" in the hovercraft is not into a world where death has been overcome by a miraculous divine love, rather, he has been saved by an earthly intervention — a sort of tender CPR — quite within the bounds of physics and chemistry. So he still has his vulnerable body and will have to die a real death one day. What he presumably has gotten over is not death but the herd's fear of death, thereby overcoming what, according to Heidegger, is the most serious constraint that normally limits people's freedom.

But if bending the rules that are accepted by the average person just amounts to being able to bend spoons, fly, and stop bullets, it doesn't seem any kind of

salvation. Being creative must mean more than just being disruptive.[24] We are lead to expect that, in return for accepting everyday vulnerability and suffering, the people liberated by Neo will be reborn to a new and better life. But what sort of life is that? To account for why it is admirable to confront risky reality rather than remain in the safe and tranquilized Matrix whatever the quality of experience in each, we need an account of human nature, so we can understand what human beings need that the Matrix world fails to provide.

But, in our pluralistic world, there are many different cultures, each with its own understanding of human nature. Even our own culture has experienced many different worlds created by new interpretations of human nature and the natural world that changed what counted as human beings and things. What mattered in the world of Homer was to be a hero and collect beautifully crafted artifacts; in the Hebrew World one had to obey God's law and to govern all other creatures; in the Christian World, the goal was to purify one's desires and to read the text of God's world in order to know God's will; and, with Descartes and Kant, people in the Modern World became autonomous, self-controlled subjects organizing and controlling objects and their own inner lives. While now, in the Postmodern World, many people, like Cypher, are egocentric hedonists trying to get the most out of their possibilities by maximizing the quality of their private experiences, and thereby treating themselves as resources.

But doesn't this just show, as Sartre famously observed, that there is no human nature? Here Heidegger makes an important meta-move. As the history of the West suggests, our nature is to be able to open up new worlds and so to transform what is currently taken to be our nature. Perhaps human beings are essentially world disclosers. So, to determine what human beings need beyond

just breaking out of the banal, it looks like we have to turn to the Heideggerian point that what is missing in the Matrix is the possibility of going beyond conventional preprogrammed reality and opening up new worlds; not just breaking the rules of the current game but inventing new games. Nietzsche says we should "become those we are — human beings who are new, unique, incomparable, who give themselves laws, who create themselves."[25] Jesus created a new world by defining us in terms of our desires rather than our actions, and Descartes invented the inner and so helped disclose the Modern World. On a less dramatic scale, Martin Luther King Jr. opened a new world for Afro-Americans.[26] It is just such a freedom to open up new worlds that the Matrix world lacks. A sense of the limit on our possibilities is what Neo experiences as the splinter in the mind. As he says to the AI intelligences at the end of the film, “I know you are afraid...of change.”[27]

Heidegger thinks that our freedom to disclose new worlds is our special human freedom, and holds that this freedom implies that there is no fixed pre-existent set of possible worlds. Each world exists only once it is disclosed. So it makes no sense to think that a computer could be programmed with rules for producing the sensory-motor connections that would allow the creation of all possible worlds in advance of their being opened by human beings. Artificial intelligences couldn't program for such a radically open world if they wanted to. In fact, programmed creativity is an oxymoron.

If being world disclosers is our nature, that would explain why we feel a special joy when we are opening new worlds. Once we experience world disclosing, we understand why it’s better to be in the real world than in the Matrix, even if, in the world of the Matrix, one can enjoy steak and good wine. Real salvation comes from transcending the world foreclosing limits of the Matrix program.

What's ultimately important to us, then, is not whether most of our beliefs are true, or whether we are brave enough to face a risky reality, but whether we are locked into a world of routine, standard activities or are free to transform the world and ourselves.

If the Matricians were simply the victims of the Matrix computer program in that they had false beliefs about the causal basis of their experiences, Neo could show them that their beliefs about the causal basis of things were false and teach them to agree with Kant that the world is an appearance, but that wouldn't set the free — not as long as they saw only the possibilities that one normally sees and never experienced anxiety. Neo has to do more. He has to do the job that Heidegger thinks anxiety does: he has to show the people in the Matrix that the order they take for granted is ungrounded and so can be creatively changed. As he says, "I'm going to. show these people a world without rules and control. A world where everything is possible."

And by the end of the movie, Neo as the One (or the anti-one as Heidegger would see it), has only begun freeing the people in the Matrix from their conformism by showing them that they have the freedom to bend the rules. He has not, however, freed them *from* the Matrix by showing them how to open new worlds. But, of course, there are two more movies to come. We can hope that, before number three is over, Neo will get to Zion and lead people in disclosing a really brave new world.

Hubert Dreyfus and Stephen Dreyfus

Endnotes

1. Names in the movie are generally very well chosen. The way the word "matrix" refers both to the womb and to an array of numbers works perfectly. Likewise, Neo is both a neophyte and

the one who will renew the world. These names are so fitting one can't help looking for the aptness of the name, Morpheus, but it is hard to find. The Greek Morpheus is the god of dreams but the Morpheus in the movie is trying to wake people up. The only way to make some sense of the name is to think of the Greek god, not as the producer of dreams, but as the one who has power over dreams: both to give them and to take them away.

2. "Imagine how his heart ached ...and yet he never blinked;
his eyes might have been made of horn or iron...
He had this trick—wept, if he willed to, inwardly."
Homer, *The Odyssey*, trans. Robert Fitzgerald (New York: Vintage Classics, 1990), 360.

Of course, the Homeric Greeks must have had some sort of private feelings for Odysseus to perform this trick, but they thought the inner was rare and usually trivial. As far as we know, there is no other reference to *private* feelings in Homer. Rather, there are many *public* displays of emotions, and shared visions of gods, monsters, and future events.

3. Saint Augustine, *Confessions*, trans. R.S. Pine-Coffin (Penguin, l961), 114.

4. Letter to Gibieuf of 19 January 1642; English in *Descartes: Philosophical Letters*, trans. Anthony Kenny (Oxford University Press 1970), 123.

5. René Descartes, "Meditations on First Philosophy - Meditations VI", in *Essential Works of Descartes*, trans. Lowell Bair (New York: Bantam Books, 1961), 98.

6. Gottfried Leibniz, *The Monadology and Other Philosophical Writings* (London: Oxford University Press), 1898. A monad, according to Leibniz, is an immaterial entity lacking spatial parts, whose basic properties are a function of its inner perceptions and appetites. As Leibniz put it: A monad has no windows.

7. Immanuel Kant, *Critique of Pure Reason*, trans. Norman Kemp Smith (New York: The Humanities Press, 1950).

8. Edmund Husserl, *Cartesian Meditations: An Introduction to Phenomenology*, trans. Dorion Cairns (The Hague: Martinus Nijhoff, 1960).

9. See, Martin Heidegger, *Being and Time*, trans. J. Macquarrie & E. Robinson (New York: Harper Collins, 1962).

10. See, Maurice Merleau-Ponty, *Phenomenology of Perception*, trans. C. Smith (London: Routledge & Kegan Paul, 1962).

11. Charles Taylor, "Overcoming Epistemology," *Philosophical Arguments* (Cambridge, MA: Harvard University Press, 1995), 12. See also, Samuel Todes, *Body and World* (Cambridge, MA: M.I.T. Press, 2001).

12. René Descartes, "Dioptric," *Descartes: Philosophical Writings*, ed. and trans. Norman Kemp Smith (Modern Library, l958), 150.

13. The point has been made explicitly by John Searle: "[E]ach of us is precisely a brain in a vat; the vat is a skull and the 'messages' coming in are coming in by way of impacts on the nervous system." *Intentionality: An essay in the philosophy of mind* (Cambridge University Press, 1983), 230.

14. There are limits of course. The Matrix programmers can't give a human being a dog's body. It's also unlikely they could make a brain in a female body the causal basis of a man's body in

the Matrix world. The hormones of the body in the vat wouldn't match the physical attributes of the body in the Matrix world.

Still, a good way for the AI programmers to prevent bodies being rescued to the hovercraft would be to give each brain the experience of a radically different body (within whatever limits are imposed by biology) in the Matrix world than the body that brain is actually in. If rescued, such people would quite likely go crazy trying to reconcile the body they had experienced all their life with the alien body they found themselves in on the hovercraft.

15. Likewise, their beliefs about entities such as viruses and black holes would be true if, like empiricists, they held that theoretical entities are just convenient ways to refer to the data produced by experiments. See Bas van Frassen, *The Scientific Image* (Oxford: Clarendon Press, 1980).

16. Of course, things are really not so simple. Most people don't have beliefs about the reality of the world; they just take the world for granted. Neo has, however, been forced to raise the question, and he believes he is now facing reality. But Neo's current beliefs could still be false. He could still be a brain in a vat fed the experience of being in the hovercraft. Given the conceivability of the brain in the vat fantasy, the most we can be sure of is that our coping experience reveals that we are directly up against some boundary conditions independent of our coping with which we must get in sync in order to act, and that, therefore, our coping experience is sensitive to the causal powers of these boundary conditions. Whether these independent causal conditions have the structure of an independent physical universe discovered by science, or whether the boundary conditions *and the causal structures discovered by science* are the effect of an unknowable thing in itself that is the ground of appearances as postulated by Kant, or even whether the cause of all appearances is a computer, is something we could never know from inside our world. But Neo does know that, as in waking from a dream, his current understanding of reality supercedes and crosses out his former one.

17. Granted it's hard to resist believing in the Matrix illusion even where causality is concerned, nonetheless, Neo learns he can stop believing in it. This new understanding of reality is described by Morpheus talking to Neo near the beginning of the movie, and by Neo at the end, as like waking from a dream. But the brains in the vats are not literally dreaming. Their world is much too coherent and intersubjective to be a dream. Or, to put it another way, dreams are the result of some quirk in our internal neural wiring and full of inconsistencies, although when dreaming we don't usually notice them. They are not the result of a systematic correlation between input and output to the brain's perceptual system that is meant to reproduce the consistent coordinated experience that we have when awake. When someone from the hovercraft returns to the Matrix world, it looks like their hovercraft body goes to sleep, but they do not enter a private dream world but an alternative intersubjective world where they are normally wide awake, but in which they can also seem to dream and wake from a dream, as Neo does after the Agents take away his mouth.

18. There is one unfortunate exception to this claim. At the end of the movie, Neo catches a glimpse of the computer program behind the perceptual illusion. This is a powerful visual effect, but, if what we've been saying is right, it makes no sense. If the computer is still feeding systematic sensory-motor impulses into Neo's brain when he is plugged into the Matrix world, then he will see the world the program is producing in his visual system. What the sight of the rows of numbers is meant to do is to remind us that Neo no longer *believes in* the Matrix illusion but understands it is a program, but even so, he should continue to *see* it.

19. The Agents, who are computer programs, don't have this freedom. It might seem that Agent Smith shows his freedom and deviates from his job of maintaining order in the Matrix when he tells Morpheus how disgusted he is with the Matrix world. We think it would be

consistent with the limitations of the Agents to understand this as Smith's playing the good cop routine; trying to get Morpheus to believe Smith is on his side, so that Morpheus, in his weakened state, will give Smith the access codes for Zion, but the movie does not exploit this possibility.

20. Not to be confused with Neo as "the One" who will save people from the Matrix. For Heidegger's account of the power of the one, see his *Being and Time*, and also H. Dreyfus, *Being-in-the-World: A Commentary on Heidegger's Being and Time, Division I* (Cambridge, MA: The M.I.T Press, 1991), Chapter 8.

21. Friedrich Nietzsche, *Beyond Good and Evil: Prelude to a Philosophy of the Future*, trans. Walter Kaufman (New York: Vintage Books, 1966). # 199.

22. Colin McGinn's essay can be found here.

23. Given the kind of bodies we have: that we move forward more easily than backwards, that we can only cope with what is in front of us, that we have to balance in a gravitational field, etc., we can question to what extant such body-relative constraints can be violated in *The Matrix* if what is going on is still to make sense.

To test these limits, the filmmakers occasionally blow our minds by using a wrap-around point of view from which action looks so far from normal as to be awesomely unintelligible. At the same time, they have successfully met the challenge of discovering which body-relative invariances can be intelligibly violated and which can't. For example, in the movie, gravity can be overcome — Neo can fly — but he can't see equally in all directions, cope equally in all directions, nor can he be in several places at once. What would it look like for a single person to surround somebody?

Time too has a body-relative structure that can't be violated with impunity. The way we experience time as moving into the future and leaving the past behind depends on the way our forward directed body leads us to approach objects and then pass them by. (See Todes, *Body and World*). Could we make sense of a scene in which someone attacked an enemy not just from behind, but from the past? If, in the movie, the liberated ones were free of all bodily constraints governing their action we couldn't make sense of what they were doing and neither could they. They wouldn't be liberated but would be bewildered, as we often are in our dreams.

24. Although being disruptive is the best one can do in the Matrix world. That's why Neo, a hacker who, as Agent Smith says, has broken every rule in the book, is the natural candidate for savior.

25. F. Nietzsche, *The Gay Science*, (Vintage Books Edition, March 1974), # 335.

26. See Charles Spinosa, Fernando Flores, and Hubert Dreyfus, *Disclosing New Worlds: Entrepreneurship, Democratic Action, and the Cultivation of Solidarity*, (Cambridge, MA: MIT Press, 1997).

27.Early in the film. Morpheus says:"What is the Matrix? Control. The Matrix is a computer-generated dream world, built to keep us under control." and James Pryor, at the end of his essay, tried heroically to make sense of this claim by speculating on what the AI programmers might do to the Matrix dwellers. If the machines had done any of these things, Pryor would have the right to say as he does: "In the movie, humans are all slaves. They're not in charge of their own lives. They may be contented slaves, unaware of their chains, but they're slaves nonetheless. They have only a very limited ability to shape their own futures. [...] The worst thing about living in the Matrix would not be something metaphysical or epistemological. The worst thing would be something political. It would be the fact that *you're a slave*."

But I fear that Morpheus is simply mistaken, at least concerning what has happened in the *Matrix* series thus far. If you're a slave, there must be a master who controls what you *can* do or, in *Brave New World*, who even controls what you *want* to do, and, of course, if you knew you were in such a world you would want your freedom. Having their causal basis used as a battery, however, doesn't interact with the Matricians' psychic lives and doesn't limit what they can decide, what they can desire, or what they can do. Pryor rightly points out, that the AI intelligences could sabotage the Matricians' projects or reset their world back to 1980 if they so chose, but what Morpheus doesn't understand (and Pryor doesn't bring out) is that there is nothing in having your causal basis used as a battery that is essentially enslaving. That is, although the Matricians' causal basis is being *used* to generate electricity, they are not being *controlled*. Their "enslavement" in the Matrix is like our relation to our selfish genes, and no one feels there is something morally wrong with our world because our DNA is using us to propagate itself; likewise the simple fact that the bodies the Matricians are linked to are serving some purpose outside their lives can't be what's wrong with living in the Matrix.

There is, indeed, a very subtle way that the AI computers have foreclosed the Matrix dweller's future but it is not by limiting the possibilities available to them *in their world*. The limitation in question has nothing to do with being brains in vats as long as the inputs to the brains are modeled on the way things normally behave in the world, and the outputs depend on the Matrix dwellers' decisions. The problem isn't epistemological, nor metaphysical, nor (pace Morpheus and Pryor) political. The problem is what Heidegger would call *ontological*. It has to do not with freedom to choose in the current world, but freedom to change worlds. By suppressing all unconventional behavior in their fear of change, and, in any case, having no way to introduce radical freedom into their programs, the AI intelligences have suppressed the Martricians' most essential human capacity - a way of being the computers can't understand but dimly fear -- our ontological capacity for opening radically new worlds.

Never The Twain Shall Meet: Reflections on The First Matrix

Richard Hanley

> 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.
> Agent Smith, to Morpheus

> And God shall wipe away all tears from their eyes; and there shall be no more death, neither sorrow, nor crying, neither shall there be any more pain: for the former things are passed away.
> Revelation 21:4, King James Bible

> Hell is—*other people*.
> Garcin, in Sartre's *No Exit*

> To deny our own impulses is to deny the very thing that makes us human.
> Mouse, to Neo

Cypher *chooses* the Matrix, and just maybe, he's not so crazy. If real life prospects are dim, then even an apparently sub-optimal alternative like the Matrix might in fact be better, all things considered.[1] But what is the *best* sort of existence for individuals like you and me? Philosophy and religion both have attempted to answer this question, and I think *The Matrix* gives us an interesting way to frame it. Is some possible "real" existence better than any possible Matrix? Or is some possible Matrix better than any possible reality? With Mark Twain's help, I shall present an argument that one important notion of the best existence, the Christian one, Heaven is after all a Matrix. The point of my polemical approach is not so much to criticize Christianity, but rather to bring the issue of the nature of ultimate value into sharper focus.

What is the Matrix? Morpheus tells Neo it's a "computer-generated dreamworld," and a "neural, interactive simulation"; it is, in other words, a

virtual environment.[2] Agent Smith assures Cypher that he won't know he's in the Matrix when he returns permanently, and it will simplify exposition to suppose that this is a necessary feature of a Matrix, while being computer-generated is not. The Matrix depicted is a mixed case, since the cognoscenti can enter it without being deceived into thinking it is real. Let us stipulate that in a *pure* Matrix, everyone is benighted, believing it is the "real deal." In most of what follows, I'll be concentrating on pure Matrices (and in the case of the Matrix depicted, on the condition of the benighted). Since we'll be discussing different kinds of Matrix, we need a name for the one depicted in *The Matrix*; Agent Smith refers to a First Matrix, so let's call the one we see the *Second Matrix*.

A Matrix, then, is an interactive virtual environment involving *systematic global deception*. Still, there are two levels of "interactivity" in a virtual environment. *Virtual interactivity* is the extent to which the environment allows, and responds to, your input. Current virtual environments are not very interactive in this sense, but the Second Matrix is. That's what makes it seem so real, at least to the benighted. (For the cognoscenti the Second Matrix it is too virtually interactive, too controllable, to seem real—at least compared with the more law-like external world.) *Real interactivity* is the potential for interaction *with others also engaged in virtual interaction*, and *real interaction* is the extent to which this potential is realized. Compare two kinds of possible Matrix: the Second Matrix is *communal*, featuring real interaction between human beings—call this *human interaction*; a *solitary* Matrix lacks human interaction altogether.

Communal Matrices differ in degree of human interaction. In the Second Matrix, billions of humans share the environment, and if we ignore Agents, it is

fully communal—every virtual human in the Matrix is an *avatar*, a virtual persona of a real human being. In the Matrix training program created by Mouse, on the other hand, virtual humans like the woman in the red dress are *simulacra*, not avatars, and human interaction during the sequence we see is limited to that between Neo and Morpheus.[3] On yet another hand, the fully communal Construct (loading program), where Morpheus and Neo watch TV, has no other virtual humans in it to interact with—and unlike the training program, it's not "big" enough to be very world-like. Call a fully communal Matrix that is big enough to be world-like, and has *many* human participants, so that human interaction is nearly inevitable, a *teeming* Matrix. (The Second Matrix is all but teeming. If we removed the cognoscenti, there would be no need for Agents, and it would be teeming.)

Now we can compare three possibilities (obviously not exhaustive) for human existence, assuming that it involves physical embodiment. *One* is the real deal, populated by other human beings: for instance, if you subjectively experience having sexual intercourse with another human being, another individual human being shares that intercourse, from another subjective point of view, because you really have physical, sexual intercourse with them. The same goes for non-sexual intercourse. If I were to meet Mark Twain (through the time travel he wrote about, perhaps), then Twain and I both would have an experience of meeting, and we really would meet, physically and psychologically. *Two* is a teeming Matrix: if you experience having (intraspecies!) sexual intercourse, another Matrix-bound human shares that intercourse, from another subjective point of view. There's no physical intercourse, of course, but there is psychological intercourse. If I have the experience of meeting Twain, then he (or some other human being) has the experience of meeting me-meeting-Twain, and there is at least a meeting of minds. *Three* is an *apparently*

teeming, solitary Matrix: if you experience having sexual intercourse, no other human is having an interactive sexual experience with you—it is like taking up Mouse's invitation to enjoy the woman in the red dress, except that you won't know "she" is a simulacrum. If I experience meeting Twain, then there is no intercourse with another human being, and neither Twain nor any other human being need have the experience of meeting me-meeting-Twain.

Our ordinary intuition is that there's something valuable about the real deal that is missing in a Matrix. Consider your present situation. You are either right now in a Matrix, thinking that it's a certain time and place when it really isn't, that a certain sequence of physical events is occurring when it really isn't, and so on; or you aren't, and it really is that time and place, and so on. Most of us hope we are *not* in a Matrix right now, which shows that, other things being equal (that is, where the experiences are identical in subjective character), we prefer the real deal. My hunch is that you also hope that, if your present existence is not the real deal, it's at least participation in a *teeming* Matrix. Being in the real deal has two distinct features of apparent value: your beliefs are more connected to the truth, and you really interact with other human beings. A teeming Matrix has less connection with truth than the real deal, but has more than a solitary Matrix, and it still provides substantial interaction with other human beings.[4] In the case of sex, there's a good sense in which you really did have sex with that other person, though in ignorance of the whole truth.[5]

If connection with truth matters so much to us, why not have the best of both kinds of existence—why not have a virtual environment, *without* all the deception? Cypher can (and does) go back temporarily into the Matrix, knowing what it is, and retain that knowledge while he is in there. But for his

permanent stay he chooses ignorance instead, because "Ignorance is bliss." Presumably, the knowledge that he is not in the real deal would undermine his capacity to enjoy the experiences, so he can't have the best of *both* worlds.[6] Intuitively, Cypher is no different from the rest of us in this regard. For a typical man, the experience of sexual intercourse with the woman in the red dress is likely to be much more satisfying if he thinks it is the real deal. Which brings us to the First Matrix.

1. What is the First Matrix?

Agent Smith's remark in the epigraph suggests that the First Matrix was, like the Second, more or less teeming.[7] Agent Smith says about the "disaster":

> Some believe 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.

The first suggestion is fascinating. Given the deadpan delivery, it is hard to say whether it posits a deficiency in the machines that designed the Matrix, or in us—in our notion of a perfect world. On the other hand, Agent Smith's own thesis seems connected with a tradition of human thought concerning the *theistic problem of evil*. If a perfectly good God exists, why does evil exist? Why is the world full of sharp corners and other hazards? A standard answer is that evil is *necessary*—it must exist in order for certain goods to exist. For instance, it is often claimed that happiness requires suffering, though this is disputable. Even if *creatures like us* can't be maximally happy, this is a reason for not creating *us* at all, and creating more felicitously instead. And does our happiness require *so much* suffering? Looking deeper, it seems clear that

virtues like *courage* and *generosity* indeed require the existence of suffering. But vices such as *cowardice* and *cruelty* couldn't exist without suffering, either—are they necessary evils, too?

The most defensible theist answer to this question is a very subtle *No, But—* : God had a choice between creating a world with *free* beings in it, or not. This choice is easy, since free will is a surpassing good. But given *libertarian* free will, which requires *causal indeterminism*, God could not know without creating the world exactly which possible world would result.[8] God might have gotten lucky, and created a world in which all free beings had only virtues, and no vices. But this is incredibly unlikely, as is a purely vicious world, and it's no surprise that He got a mixed world, with most humans having virtues *and* vices. The picture that emerges is that a world with human beings in it is a world with sharp corners (*natural evil*) to provide genuine free choice, and so very likely contains sin (*moral evil*) as well. Call this the *Free Will Theodicy*. Its assumption that free will is *libertarian* free will—requiring causal indeterminism—is Christian orthodoxy, so I grant it for the sake of the argument.

Filling in the details of the theodicy, focus on the *will* itself. Our actions are ultimately explained by what we want, most especially by our *non-derived desires*.[9] In a world of sharp corners, not all these desires can be satisfied. Indeed, there often will be conflicts between individuals in what they desire—one person getting what they want means that another doesn't. (Presumably, God could not arrange a concordance of wills—substituting for conflicting desires, or deleting them altogether—without eliminating free will.) Indeed, the existence of other human beings in the world is *part* of the "sharp corners"—a source of suffering— in addition to being a source of moral evil.

And not just because others are in competition with you for resources—sometimes others *are* the resource, as the sexual intercourse example shows. If you badly want sex with another person and they badly don't want it with you, then someone is going to suffer.

If the Free Will Theodicy is correct, then God can only control the non-human environment. Each human being is a part of the environment of every other human being, so as soon as you put more than one creature with libertarian free will into the mix there will, absent astonishing coincidence, be tears. You can minimize the effect human beings have on each other, but only by minimizing their interaction (say, by putting each on a separate planet). Even then, as long as human beings desire interaction (as a means to things we want, such as to procreate, and perhaps even for its own sake), mere isolation won't solve the problem.

The creators of the First Matrix tried to produce a relatively good existence for Matrix-bound humans. (We needn't suppose the machines were benevolent; perhaps the bioelectric-to-fusion reaction process is more efficient the happier humans are.) In doing so, the machine creators had some of God's problems. They presumably lacked some of God's creative abilities, but they also had fewer constraints, since God is supposed to be no deceiver.[10] Why was the First Matrix a disaster? If the machines were trying to produce an existence with *no human suffering*, then perhaps they tried the wrong design: a teeming Matrix populated with otherwise typical human beings. Even if the machines removed a lot of sharp corners (no volcanic-eruption, or man-eating-shark experiences), as long as there is interaction with other human beings plugged into the same virtual environment, someone is going to suffer, as the example of sexual intercourse demonstrates. This attempt would not produce a Matrix

where "none suffered," and the suggestion fits badly with Agent Smith's remark, "No one would accept the programming." Let's discard it.

Which leaves two basic choices: the machines either substantially altered the nature of human beings in the First Matrix (say by arranging a concordance of wills), or else they created a solitary Matrix for each human being. The advantage of a solitary Matrix is that the virtual environment can be completely tailored to an individual's desires—perhaps the Matrix "reads off" the content of desires from his brain, anticipating a little, matching its programming as far as possible to the satisfaction of his desires as they develop and change.

Perhaps a battery of solitary Matrices was beyond the machines' practical resources, but let's suppose not—clearly it's in principle possible for them to have done things this way. However, if Christians are correct, and our wills are in fact undetermined, then our desires cannot be fully anticipated. There is bound to be a gap between the evolution of our desires, and the Matrix's capacity to satisfy them; hence some suffering is inevitable. This would partly explain Agent Smith's remark, but once again would not explain why "No one would accept the programming."

We are left with two possible explanations of the remark: either humans by their nature could not be successfully altered through programming; or else unaltered humans were psychologically incapable of accepting the relevant virtual environment. The latter seems to be Agent Smith's thesis: the "perfect world" was just too good to be true, and literally incredible.[11] Are we human beings simply incapable of having a happy existence, with no suffering? Not on the standard Christian view, according to which just such an existence awaits us in *Heaven*.

II. What is Heaven?

The Christian notion of Heaven is far from a settled body of doctrine. (For instance, are there literally streets paved with gold, or is this just a metaphor for some barely imaginable, wonderful state of affairs?) Nevertheless, it has been asserted with some authority that the human condition in Heaven will be very different from that here and now. It is agreed that there is no suffering (see the epigraph), not to mention "exceeding joy," (an expression which occurs four times in the King James Bible), but what exactly will we do there? Some of the common claims about this can seem puzzling. In *Letters from the Earth*, Mark Twain has the banished Satan report to his fellow angels on the beliefs of mortal Man:

> For instance, take this sample: he has imagined a heaven, and has left entirely out of it the supremest of all his delights, the one ecstasy that stands first and foremost in the heart of every individual of his race — and of ours —sexual intercourse! ...
>
> His heaven is like himself: strange, interesting, astonishing, grotesque. I give you my word, it has not a single feature in it that he actually values. It consists — utterly and entirely — of diversions which he cares next to nothing about, here in the earth, yet is quite sure he will like them in heaven. Isn't it curious? Isn't it interesting? You must not think I am exaggerating, for it is not so. I will give you details.
>
> Most men do not sing, most men cannot sing, most men will not stay when others are singing if it be continued more than two hours... In man's heaven, everybody sings! The man who did not sing on earth sings there; the man who could not sing on earth is able to do it there. The universal singing is not casual, not occasional, not relieved by intervals of quiet; it goes on, all day long, and every day, during a stretch of twelve hours. And everybody stays; whereas in the earth the place would be empty in two hours...

Satan's list is long, and frequently amusing:

> I recall to your attention the extraordinary fact with which I began. To wit, that the human being, like the immortals, naturally places sexual intercourse far and away above all other joys — yet he has left it out of his heaven! The very thought of it excites him; opportunity sets him wild; in this state he will risk life, reputation, everything – even his queer heaven itself — to make good that opportunity and ride it to the overwhelming climax. From youth to middle age all men and all women prize copulation above all other pleasures combined, yet it is actually as I have said: it is not in their heaven; prayer takes its place.

His main observations we can summarize as: (i) Man thinks he will be blissfully happy in Heaven; (ii) no activity that Man finds blissful on Earth will he pursue in Heaven; (iii) the activities that Man thinks he will pursue in Heaven are ones he avoids whenever possible, here on Earth. Call this appearance of inconsistent values, *Twain's Puzzle*. In Mouse's terms, it seems that we think we will be happiest denying our own impulses. Satan somewhat overstates the puzzle when he writes that Heaven "has not a single feature in it that [Man] *actually values*." Man thinks that in Heaven he will still value joy and disvalue suffering, for instance. Satan's point is that Man appears to think that his *desires* will be radically different in Heaven: he will desperately want the things that he does not want at all now, and not want at all the things that he desperately wants now.

Does Man think his *will* is going to be different in Heaven? That depends. *Psychological hedonism* is the view that there are really only two non-derived human desires: to obtain pleasure and avoid suffering. If this were true, then Man's will does not change if he merely changes his beliefs about what it is that will bring him pleasure and avoid pain. If psychological hedonism isn't true (and Christians seem—wisely—to think it isn't true), then a case can be made that (according to Satan, anyway) Man expects his will to be altered in Heaven.

Contrary to Satan, it can be argued that at least where sex is concerned, the Christian view is that such impulses *ought* to be denied, and the relentless pursuit of gratification is, in a Christian, a matter of *weakness* of will, not in its constitution. It might be further claimed that giving in to such impulses actually causes you suffering. This makes some sense in the case of, say, a married man tempted to adultery, whose guilt may prevent him from full enjoyment. Suppose that in Heaven, since there is no marriage (so says Jesus, see for instance Matthew 22:30), there is really no one psychologically "safe" to have sexual intercourse with, and you would inevitably feel guilty about engaging in it. Then the elimination of suffering requires the elimination of sex. (Of course, Satan and Mouse would no doubt respond, with some justification, that this is all premised on the belief that sex outside marriage is something bad in and of itself, a notion you happily will be disabused of in Heaven. But the question is what the typical Christian believes, whether it is true or not.)

Leaving aside what you would do there, believers in Christian Heaven commonly hold the following four theses about it:

(1) It's possible for a human being to be in Heaven. More precisely, if all goes well it will be you that survives bodily death and goes to Heaven.
(2) Human beings in Heaven will experience happiness, but no unhappiness.
(3) Human beings in Heaven possess free will.
(4) Human beings in Heaven interact with other human beings in Heaven.

It's worth expanding on (1). Christians standardly expect to recognize their

loved ones in Heaven, which presumably requires remembering them.[12] So it seems that they expect considerable psychological continuity between their Earthly and Heavenly existences—perhaps this is even guaranteed by the requirement that God be no deceiver. But such psychological continuity sits uncomfortably with (2). Christians on Earth are typically saddened by the fact that unbelievers will not get into Heaven. It seems that, if anything, they would be sadder still, when confronted by the wonders of Heaven, knowing that the unsaved are residing instead in "the lake which burneth with fire and brimstone." And it would seem to be cause for special anguish if one of your *loved ones* is absent from Heaven. (Another version of the problem arises with *missing* your loved ones—being sad, not for them, but for yourself, that they are not around. Even if you don't miss sex with your Earthly spouse, it seems you would miss *them*.)

Heaven is also widely supposed to provide an opportunity to meet human beings you never knew on Earth. But if I'm in Heaven, and I really want to meet Twain, then I will be sadly disappointed if he isn't there (and angry, if it's all on account of those *Letters*). Moreover, certain truths will presumably be available to you in Heaven. Suppose that Mother Theresa is your idol, and you can't wait to tell her so. However, you find out she's not really a saint—indeed—quite the opposite, and not in Heaven at all. You may be upset not only for your own sake, but for the sake of humanity (you may respond with a quite cynical attitude toward human nature). Heaven seems on the face of it to provide many opportunities for suffering.

There are three basic ways around this sort of problem. First, suppose *universalism*—the doctrine that everyone gets into Heaven—is true. This will solve the problem only if, upon entering Heaven, Christians no longer believe

that there ought to be any qualification for it (else they likely will be annoyed that others got a "free pass," especially a holier-than-thou like Mother Theresa). Second, God could suppress the knowledge that others are not in Heaven. But this requires Matrix-like deception (either to provide the appropriate virtual interaction with non-avatars, or else to just delete all memory of the missing), and Heaven would not be the real deal. Third, perhaps what we care about—our desires—will change, so that good Christians no longer will mind the fact that others—even loved ones—are suffering (they might even take pleasure in it). But to accept this raises an acute version of Twain's Puzzle.

All in all, it may be better to revise (2) to:

(2*) Human beings in Heaven will be as happy as they can possibly be.

We may thus grant that it's not possible for *all* suffering to be absent in Heaven—though this requires taking *Revelation* less literally than many Christians do.

(4) is taken completely for granted, as far as I can tell. Part of the point of Heaven is to be reunited with (saved) loved ones, and to engage in "fellowship" with the other inhabitants. But what of (3)? According to the Free Will Theodicy, free will is a surpassing good, so on the face of it, Heaven *must* include free will. Yet Heaven is a place without sin. And according to the Free Will Theodicy, sin is to be explained by the presence of free will in the world. To deny (3) also raises Twain's puzzle. We believe we now have libertarian free will, strongly desire it now, and are devastated at the thought of losing it. If God is no deceiver, then if (3) is false, we would in Heaven *know* that we have no free will. Yet, presumably, we would not mind—be blissful, yet not ignorant.

Like the builders of the First Matrix, God has two main choices in creating a Heaven for human beings: either substantially alter the nature of human beings in Heaven (say by arranging a concordance of wills, contrary to (3), and perhaps even contrary to (1)), or else put each human in a solitary Matrix, contrary to (4). One advantage for denying (4) is that (2) has the best chance of being true, as long as the solitary Matrix provides plenty of (virtual) interaction with virtual humans. Those in such a solitary Matrix will think they are in the real deal. They'll think they are in Heaven, along with everyone that they want to be there, and nobody that they don't want there. They will think they get along with along with everyone else just fine; that there's no sadness, no sin, and so on. God knows what they freely want, and tailors each virtual environment to provide exactly that, if possible. (If it's not possible, because they freely want to be in the real deal, this lack is not *experienced*, and so is not a source of suffering.)

Just as it did with the First Matrix, libertarianism raises a difficulty, since you might think that God could not know what you want, when this is undetermined. Some medieval Christians resolved the problem of the compatibility of free will with God's foreknowledge by supposing that changeless, omnipresent God knows the (causally undetermined) future by, so to speak, already having been present then, and having seen what happens. God knows what you do *because you do it*, and not vice versa, hence you may do it freely. The same resolution can be applied here, as long as time exists in Heaven: God knows what you *will* want before you want it, by having been in the future and (so to speak) looking into your mind then.[13]

Can (3) and (4) both be maintained, given (1) and (2*)? There is logical space

for this possibility. (3) can be true, and yet there be no sin in Heaven, *if* Heaven is like the lucky roll of the creation die: the world where free beings always choose rightly. In Heaven, everyone will be free to sin, but just *doesn't*. The immediate problem with this suggestion is that it seems incredible that such a coincidence will actually obtain. Perhaps we can appeal to a difference between this situation and that of creation: God has a chance to observe the behavior of free individuals, and only admits the deserving—those who actually don't sin while on Earth. But this would get hardly anyone into Heaven. Worse, it seems to give inductive support, but no *guarantee* at all, that unblemished individuals won't sin ever in the eternity they spend in Heaven.

It is standardly claimed that all are free to sin in Heaven, but none do, because they are in some sense *incapable* of doing so; no one can sin when they are at last with God. This raises two distinct problems. The first is that any such incapability seems incompatible with libertarian freedom, rendering (3) false after all. The second is that, if there is no incompatibility between human beings having libertarian free will and being incapable of sin, then the Free Will Theodicy seems to collapse. God could have just created Heaven and be done with it, a creation with all of the benefits and none of the disadvantages.

In addition to the problem of sin, we might wonder how it can be managed that free human beings, all interacting with each other, have no desires in conflict. As Satan observed, it must be that our desires change radically. But what ensures this? If it is inevitable that they change in this way, then libertarian freedom is again threatened. And if we are somehow free anyway, when our desires are radically altered, then why didn't God just turn this trick to begin with, and spare all the lost souls? Perhaps we should also consider Mouse's point. If our desires change too radically, will we still be *human*

beings, as (1) would have it?

III. Conclusion

Perhaps both explanations of the failure of the First Matrix are correct. Recall the suggestion that machines could not program our "perfect world." Perhaps our thinking is incoherent: we think that the best existence is one where human beings interact with each other *and* everyone has libertarian free will *and* nobody suffers *and* that someone knowingly arranges this. If this is an incoherent notion, not even God can actualize it.

In creating the Second Matrix, the machines went for interaction combined with free will (which we are assuming is libertarian), with the overwhelming likelihood (inevitability, in practice) of suffering. We can now explain Agent Smith's remarks: if we rank the elements of our incoherent notion of the best existence, human interaction and libertarian free will rank above the absence of suffering. And since they jointly require (*almost* by definition) the presence of suffering, it can be said more or less truly that we "define [even the best] reality through misery and suffering." The First Matrix was an attempt to give interacting humans an existence free of suffering, but this program *required* a radical revision in their wills, contrary to libertarian free will, and so "no one would accept the program." Mouse might say it was an attempt to deny the very nature of human beings.

If the real deal includes libertarian free will, then so does the Second Matrix—our desires, though often enough unsatisfied, will be after all undetermined. (The sense in which humans are liberated from the Matrix has nothing to do with libertarian free will, which can be enjoyed behind bars.) The Second Matrix also features substantial variation in wills amongst its human

inhabitants, and the interesting ethical choices that arise when this is so. For example, apart from the Agents, each virtual human is an avatar, and the "good guys" in the movie end up killing a lot of human beings in their fight against the Agents. It's hard to view these human beings as collaborators, given the nature of the Matrix, so their deaths presumably are to be regarded as acceptable collateral damage, inevitable given the difference in desired outcome. All in all, the Second Matrix is the machines' best attempt at matching what Christians believe God did for us through creation. [14]

When we humans turn our eyes toward Heaven, our ranking of values seems to change, and Twain's Puzzle arises anew. In Heaven, there is a heavier weighting given to the absence of suffering. God can knowingly minimize suffering in a real deal, while retaining human interaction, but at the cost of libertarian free will. But given that Heaven is supposed to involve no suffering at all, and given the surpassing value of libertarian free will in the Christian view of things, God's choice is clear: Heaven is a solitary Matrix.[15] The machines, not being God, did not know that *Heaven is no other people*. Never the twain—Twain and I—shall meet (in Heaven, anyway—there's always the *lake*, I suppose.)

A relative of Twain's puzzle emerges. We when consider a pre-Heaven existence, we seem to prefer the best real deal to the best Matrix. When thinking about Heaven, we seem to prefer the best Matrix to the best real deal. This schism in our thinking is represented by the two competing visions in *The Matrix*: on the one hand is the Matrix, and on the other is Zion—named ironically, if I am right, for God's Holy City in Heaven—the place in the bowels of the Earth where human beings not in the Matrix dwell.

Endnotes

1. See Christopher Grau's essay, "The Experience Machine." Indeed, I recommend you read Grau's essay in its entirety before proceeding.

2. Metaphysicians will not yet be satisfied. "Matrix," is from the Latin for "mother," and originally meant "womb" (it is used in the Old Testament five times with this meaning), or "pregnant female." In several contexts it means a sort of substrate in which things are grown and developed. Given this etymology, the Matrix might have been the concrete thing that includes the collection of deceived humans in their vats. A more modern meaning of "matrix" is based in mathematics: a rectangular arrangement of symbols. Perhaps "the Matrix" (an expression surely borrowed from William Gibson's earlier use in Neuromancer) denotes the array of symbols encoding the virtual environment, which we might distinguish from the environment itself. But *The Matrix* gives the impression that the environment just is the array of symbols that Neo sees when he finally sees in—so to speak—Matrixvision. Its concrete-world-like appearance seems an inferior perception. (The Matrix thus seems allegorical in turn of Plato's well-known allegory of the Cave; Neo is enlightened about his own nature by liberation from the Matrix, and by the end he sees the true nature of the Matrix.) Still, it is the concrete-world-like appearance of things that I'm concerned with here, so let's ignore the possibility of a Neo.

3. I use *simulacrum* in the following sense: "something having merely the form or appearance of a certain thing, without possessing its substance or proper qualities; a mere image, a specious imitation or likeness, of something." (OED) It is also a nod towards Baudrillard, whose work Simulacra and Simulation both influences and appears in *The Matrix*. See my essay, "Baudrillard and The Matrix."

4. Here's an interesting question: which is better, the Second Matrix, or a systematically deceptive personalized non-virtual environment—a *Truman show*—that you never discover the true nature of? The latter has more veridical human interaction in one sense, because you really physically interact; but the interaction is less veridical in another sense, in that other human beings are willing participants in the deception. Another case to think about is a solitary Matrix allowing interaction with non-human participants (dogs, perhaps). Another still is a solitary Matrix without even the appearance of real interaction — call this a lonely Matrix. I don't know about you, but I prefer Sartre's vision of Hell to a lonely Matrix.

5. The Second Matrix may connect with the truth in some unnecessary ways. One's virtual body is depicted as more or less veridical, for instance. (But this may be only "residual self image," as Morpheus tells Neo. If Cypher were put back into the Matrix as Ronald Reagan, that would be clinching evidence that one's avatar can be strikingly different.) Breaking this connection would permit interestingly different human interaction: for instance, you could unknowingly have an experience of heterosexual intercourse with another (unknowing) human who is in fact of the same sex.

6. Sometimes it is argued that you are better off—happier—being a Christian, even if God does not exist. If Christian belief is easier to maintain inside the Second Matrix than outside it, then Cypher could have an extra pragmatic reason for going back in.

7. Is Agent Smith telling the truth? I have no idea. He is attempting to "hack into" Morpheus's mind to gain the access codes to the Zion mainframe computer, so in interpreting the story we

should take everything he says—and so, even the very existence of the First Matrix—with a grain of salt. For my purposes, though, we can pretend that he's telling the truth.

8. We need not fully characterize libertarian free will for present purposes. The main point is that causal indeterminism is a necessary condition of it. Causal indeterminism is the denial of causal *determinism*: the thesis that every event is completely determined by causally prior events. A useful and common illustration is to ask whether or not everything that happens, or will happen, is in principle *predictable* — this will be so if determinism is true, and not so if indeterminism is true. (Whether the future can be known by means other than prediction is a different question — see note 11.) The thesis that we have libertarian free will is called *libertarianism*.

9. Many of our desires are *derived* from other desires plus belief, for instance if Ralph desires to kiss Grandma only because he desires an inheritance and he believes kissing Grandma is necessary to achieve this. Non-derived desires, such as Ralph's desire to kiss the girl next door, are importantly independent of belief—they are had, so to speak, for their own sake—and seem to constitute what we refer to by "the will."

10. Is this a theological guarantee of the real deal? The Christian can surely deny this. The existence of the Matrix seems compatible with God's being no deceiver, given the Free Will Theodicy, if the machines have libertarian free will. And if they do not have libertarian free will, as long as they are the product of human free will, they are not part of the environment God knowingly created.

11. I am reminded of a passage in William Gibson's *Count Zero*: "Eyes open, he pulled the thing from his socket and held it, his palm slick with sweat. It was like waking from a nightmare. Not a screamer, where impacted fears took on simple, terrible shapes, but the sort of dream, infinitely more disturbing, where everything is perfectly and horribly normal, and where everything is utterly wrong."

12. People seem to expect that their body in Heaven will resemble their Earthly one (just as their Matrix "body" seems to resemble their real one). Perhaps this is for purposes of recognition, but it seems unnecessary—common memory can do the job.

13. It would be intriguing if God could "cheat" by doing what he does *because* He sees, from the way the future is, what He will do. This would raise a fatalist, *bake-your-noodle* puzzle like the one the Oracle raises for Neo's smashing of the vase. But God is a special case. Being unchanging, He cannot be *caused* to act on the basis of future knowledge, and there is little metaphysical sense to be made of "He did it because He did it."

14. The typical Christian is a *Cartesian dualist*, believing they are a spirit or soul distinct from their physical body, and that embodiment provides the means for human interaction. Loosely speaking, then, our physical bodies are the "avatars" of the real us, in a more or less "teeming" physical environment. The Second Matrix is in this respect almost the converse of Christian creation.

15. Perhaps Christians have had this revelation available to them all along. Luke 10:20 has Jesus telling his disciples, "... rejoice, because your names are written in heaven." In Latin, "matrix" also meant a list or register of names (also, *matricula*, hence our English verb *matriculate*). Intended meaning can go astray: according to some, the notion that the fruit of the tree of knowledge of good and evil was an apple, rests on a confusion over the Latin malum, meaning both "evil" and "apple tree." In like manner, maybe Jesus's message, lost in translation, was that Heaven is a Matrix!

16. "People" in the sense of *human beings*. It might be objected that there has to be at least

one person you are in contact with: God. I'll just concede this, since it doesn't affect the argument, God's desires presumably being compatible with yours. (Real interaction with angels likewise presents no problems.) A fascinating further suggestion is that you couldn't be maximally happy unless the "program" was extremely sophisticated, and then it might be objected that we should regard the solitary Matrix as containing *virtual* individuals—such as your imaginative sexual partner(s), if there is sex in Heaven—which are arguably persons you really interact with. (Agent Smith's impassioned outburst that he *hates* the Second Matrix might be evidence of personhood, for instance.) *If* these virtual individuals are persons with libertarian free will, then you can't interact with them either, without someone eventually suffering. So we might have another argument that the Christian Heaven is an incoherent notion.

Reality, What Matters, and *The Matrix*

Iakovos Vasiliou

The Matrix is, at its core, a film with a moral plot. We, the viewers, like the heroes, are in on a secret: The reality that forms the lives of millions of human beings is not real. The world that seems real to most people is in fact a computer-generated simulation, but almost no one knows it. In reality human beings are floating in liquid in machine pods, with tubes connected to them in a grotesque post-apocalyptic world where the sun is blotted out. To the average person, of course, it seems to be the ordinary world of 1999. Although some details of the history remain untold, it is an essential part of *The Matrix* that we are provided with a specific account of how all of this happened. There was a battle between human beings and machines whose cognitive capacity had surpassed their own. In a desperate attempt to win, human beings blocked out the sun's light in order to deprive the machines of their power source. Despite this extreme tactic, the humans lost, were enslaved, and are now farmed to supply energy sources for the machines. The machines induce the appearance of ordinary 1999 life in the human beings with a computer generated "virtual community" for the purpose of keeping them docile and asleep so that they and their offspring can be used like living batteries. While humans seem to walk around in an ordinary life, their minds are radically deceived and their bodies are exploited. The heroes are thus depicted as fighting a noble battle for the liberation of the human species.[1]

I have so far drawn out two aspects of the "moral background" of the film: enslavement and deception. We should also note the perspective we have on

the Matrix as viewers of *The Matrix*. We have what is sometimes called a "God's eye" perspective: we can see *both* the Matrix reality *and* "real" reality. We are let in on the truth about the situation, and we are not supposed to question, for example, whether the battle between Morpheus and his friends and the Agents is itself being conducted in another "meta-matrix", or whether the view of the human pods we see might only be some sort of dream image or illusion. As viewers of the Matrix, we are in on the truth and we can see for ourselves that human beings are both enslaved and deceived. Given the outlined history, we are meant to understand the situation of the humans as a terrible and unfair one.

1. How Does the Matrix Differ from Reality?

Excluding, for the moment, the heroes – Morpheus, Trinity, eventually Neo, and the rest of their crew – and the machines, no one *in* the Matrix shares our God's eye perspective. In everyday life as well, as far as we know, reality is simply there. When we watch the film, we identify with the heroes in part because we are repulsed by the idea that human beings are enslaved and deceived.[2] It is easy to find these two elements at work in *The Matrix* in part because we think of enslavement and deception as things that are done to some people by others; one group of people enslaves another, or one person or group deceives others. In the film it is the machines who are the agents of slavery and deception and almost all of the humans are victims. But how does the Matrix, and the situation of the ordinary people within it differ from reality and the people within *it* (i.e., us)?

Let's begin with enslavement. We are forced to do many things in ordinary reality: we must eat, drink, sleep, on penalty of death. Also, no matter what

we do, we shall eventually, within a fairly predictable time frame, die; we cannot stay alive forever, or even for a couple of hundred years. We can't travel back and forth in time; can't fly to other planets by flapping our arms. The list could go on and on, and I have simply offered limits we are subject to in virtue of the laws of nature. In other words, compared with some easily imaginable possibilities, we are severely constrained, in a type of bondage, though ordinarily most of us don't think of it as such. Writers, artists, philosophers, and theologians over the centuries have of course been keenly aware of these limitations, examined many forms of human bondage, and offered various types of suggestions as to how to free ourselves. Human beings have longed to "break out" of this reality, to transcend the imposed limitations on their physical being. Moreover, we should be clear that these limitations are *imposed* on us. We simply find ourselves in this condition, with these rules: we all die within approximately 100 years. It has nothing to do with our voluntary choice, our wishes, or our judgements about what ought to be the case.

Who has done this to us? Answering this question is important to some degree because we typically use the term "enslavement" to refer to something done by one agent to some others. In the case of the constraints I outlined above it may be harder, initially, to find anyone on whom to pin the blame. But of course human beings have offered answers to this question: one is God; another, the laws of nature. Religious thinkers have struggled with questions about why we should not be angry at God for constraining us in the ways he does: why do people die, why can't we go back in time, travel to other planets, etc.? Others conclude that God is not constraining us, but simply the laws of nature. At least at first this thought might be a bit more palatable insofar as we think of the laws of nature as *impersonal* features of reality; no one made them that way (if God did, then we get angry at him again). They do not *mean*

to constrain us and there is no mind or intelligent force actively doing anything to us.[3] Either way, however, our actual situation is one of involuntary constraint, much akin to the humans' situation in the Matrix, except that it is not at the hands of machines against whom we lost a war, but at the hands of God or "nature".

The second aspect of the moral background of *The Matrix* is deception. Human beings are being actively deceived by the Matrix into believing things about reality that are not true. Deception offends many people, except perhaps for committed subjectivists, since many people believe that they want to know, or at least have the right to know, the truth, even if it is terrible. For one person, or a group of people, purposefully to keep others in the dark about some truth is to diminish the respect and authority of those people; it is to act patronizingly and paternalistically. In such situations, a few people decide which truths others can handle, and which they can't. Although this happens routinely – consider the relationship between those who govern and those who are governed – many people bristle at this idea and want the scope of such filtering of the truth to be severely limited.

We might think, however, not about the deception of some people by others (just as we did not look at the enslavement of some people by others), but the deception of humanity in general. In Homer's *Iliad* and *Odyssey* the gods are depicted throughout as capriciously deceiving human beings, compelling them knowingly and unknowingly to do specific things, and generally interfering quite frequently in human affairs. The humans in Homer certainly seem to be caught in a matrix of sorts, with gods and goddesses operating on a plane of reality that is not accessible to them (unless the gods want it to be) but that nevertheless often affects matters in the humans' ordinary reality. As human

beings began to understand that the Earth rotated around the Sun, and not vice versa, Descartes certainly worried about the extent to which God had had a hand in deceiving all of humanity for tens of thousands of years up to that point. He devotes a significant portion of the *Meditations*[4] to worrying about how an all-good, all-knowing, and all-powerful God could have allowed (and whether indeed he was complicit in) people's radical deception about the relative motions of the planet they live on, and other truths that turn out to be radically different from how things *seem* to be.

So in our ordinary situation, without any cruel machines doing anything to us, we realize that there are nevertheless many things we cannot do, and we know that we humans have been radically deceived by natural phenomena (or by the gods, or by God) about things in the past, and that it only stands to reason that we may be radically mistaken about our explanations of things now. I say people's "radical deception", despite the fact that, as with being enslaved, being deceived also seems to require an agent – someone to *do* the deceiving. We should note, however, that we talk of being deceived or fooled by mirrors, or by the light, or by angles. Natural phenomena are often described as contributing to our misunderstanding of them for a reason. Even though human beings were mistaken for millennia about the fact that the Earth moves relative to the Sun, and not the other way round, it is hard to describe our error as simply having "made a mistake", as though humanity forgot to carry the two in some addition calculation. Surely part of the reason that it took humans so long to understand the motions of the Earth is that the appearances themselves are deceptive: it certainly *looks* as though the Sun is moving across the sky.[5] We can see the very development of philosophy, art, religion, science, and technology as all stemming from a drive to "free humanity" from

such deception and enslavement, as part of a struggle to achieve the position of a Morpheus or a Neo.[6] We develop planes to break the bonds of gravity that keep us physically on the surface of the Earth; we develop complex experiments and gadgets designed to discover the truth about things independently of how they may appear.

My first point, then, is that if we could get hold of the being responsible for setting up the reality we're actually in, then we could perhaps "free" ourselves, finally knowing the full truth about things, and being able to manipulate reality. If God is responsible, we would need to plead with him successfully, or to fight him and win; if it's the mathematical formulae (computer programs?) underlying "the laws of nature," we would need to learn how to write and rewrite them. We would then all be Neos.[7] We might note too, at this big-picture level, a difference between the Homeric gods and the Judaeo-Christian-Islamic God. In Homer's world the gods were frequently literally in battle with humans who were greatly outmatched, although not entirely impotent – much like the humans that, before Neo, fought with the "Agents". With the God of the major contemporary religions, he is, by definition, all-good. From this perspective, we should not *fight* God, for he set things up the way he did for a wise and benevolent reason; rather, we need to learn to accept the position he has put us in (this "mortal coil", our reality, our matrix) and, then, if we act certain ways, or do certain things, he will free us from this reality after we "die" (i.e. not go out of existence, but end our stay in this reality) and show us the truth in heaven.

I hope this necessarily brief discussion enables us to see the importance of both the God's eye perspective and the moral background of the film for effecting a difference between the situation depicted in *The Matrix* and our

ordinary condition. As viewers of the film we are in a special position: we can see both inside and outside of the Matrix. We can see that it is not a benevolent God who has set up this 1999 reality, replete with constraints and deceptive appearances, pain and toil, for some wonderful, miraculous purpose. Nor is the reality of most people in the Matrix the result of impersonal laws of nature. Instead, machines who use human beings as batteries are responsible for what counts as reality for most people. *The Matrix* then supplies us the viewers with a definitive answer about who is responsible for what most human beings take to be reality.[8]

2. A Benevolently Generated Matrix

Now *The Matrix* could be significantly altered, without changing anything in the Matrix. Imagine that the real world is a post-apocalyptic hell, just as in the film, but, unlike in the film, suppose that the cause of the world's being in such a state is not some battle with machines that wanted to enslave us, but the emission of so many greenhouse gases with our three-lane-wide SUVs that we completely obliterated the ozone layer and thereby rendered the planet uninhabitable by us or by the plants and animals that we rely on for our survival. Suppose further that sometime in the future, in order to *save* the human race, scientists set up an enormous self-sustaining machine, just as in the film (minus the scary "Sentinels"), designed to keep the human species alive and reproducing for the 100,000 years it will take for whatever weeds are left on the planet to fix our atmosphere and make the planet once again habitable in a normal way. The machine operates simply on solar power (since, on this scenario, the sun is now stronger than ever, frying almost everything else on the planet), so that human beings are not needed as "batteries".[9] While humans are stuck in this state, the scientists create the Matrix for them

to "live" their lives in instead of being conscious of floating in a vat for the length of their life, which would clearly be a most horrific torture. Once the power of the sun is diminished to a habitable degree (because of the repaired atmosphere) the machine would "wake" us humans, and we could go back to living on the planet.

The ordinary person in this scenario is in the same condition as an ordinary person in the film, except that instead of the Matrix being the diabolical result of evil machines who exploit the human race, it is the result of benevolent human beings trying to keep the human race alive in as good condition as possible under the terrible circumstances. Of course it would *seem* no different to the person *in* the Matrix. We, the viewers, however, would have quite a different response to *The Matrix*. There would be no enemy to fight, no injustice to rectify (the pushers of SUVs being long dead). If there were a Morpheus in this situation, how would we think of him? If Morpheus and his friends had left the Matrix, and figured out that they could, with extreme difficulty, survive in the devastated world (eating disgusting porridge, etc.), should they go about "freeing" everyone, even if it would take another 10,000 years for the Earth to return to its present state of habitability?

As Chris Grau discusses in his introductory essay (section "C"), the Matrix is importantly different from Robert Nozick's "experience machine".[10] Grau points out that we retain free will in the Matrix. The "world" in the Matrix will respond to our free choices, just as the ordinary world does now. Another difference that I think is quite significant is that in the Matrix, unlike in the experience machine, *I am really interacting with other human minds*. There is a community of human beings. With the experience machine, it is all about *my* experience, which is the private content of my own consciousness. It is

imaginable that I am alone in the universe, floating in a vat set up by a god who has since committed suicide. In sceptical problems that stem from the Evil Genius hypothesis in Descartes' first *Meditation*, there is a threat of solipsism and the dread of feeling that one might be alone in the universe.[11] In the Matrix, however, when two people meet there are really two consciousnesses there that are each experiencing "the same things" from their respective positions. Everyone is hooked up to one and the same Matrix; there are not unique matrices generated for each individual. Of course people aren't really shaking hands – their hands are in vats – but it seems to each of their consciousnesses, not just to one consciousness, that they are shaking hands. This feature of the Matrix is also a respect in which life in the Matrix is critically *unlike* a dream, despite the fact that the humans are described as "dreaming".[12] Regardless of the amount of conscious control one has or lacks in a dream, a dream is *private* to one's own consciousness. It is part of the grammar of "dream", as Wittgenstein might say, that only I can have my dream.[13]

Now this seems to me to be of enormous significance in thinking about the Matrix. If two people fall in love in the Matrix, in what sense would their love not be real? It would not be as if a person merely *dreamt* that he had fallen in love with someone; for in a dream that person is not really there at all, just like in Nozick's experience machine. It is true that in the Matrix they would not really be giving each other flowers, or really holding hands. They would, however, both be experiencing the same things together. They would know each other as persons, who display their characters in how they react to all of the – in one sense – "unreal" situations of the Matrix. Moreover, people in the Matrix really suffer and experience pain, and when they die in the Matrix, they

die in the "real world" too. The fact that one and the same Matrix is inhabited by millions of minds means that millions of people are *really* interacting, even if the physical universe in which they are interacting is radically different from how it appears.

Consider as well writing a novel, a poem, or a philosophy paper. Or consider painting or dancing, making music or a movie. Would any of these activities be affected by the fact that what I took to be material objects were objects that were computer generated? And if not, in the benevolently generated Matrix I hypothesized we would seem clearly better off as a species, developing artistically, intellectually, loving each other within the Matrix rather than fighting for survival and barely succeeding outside of it. If my aim in life was to write some extraordinary philosophy or a ground-breaking novel, surely I could do that far better within the Matrix than outside of it where a person must battle simply for his or her survival. After all, where does my novel or my philosophy paper exist for much of its genesis and storage? In a computer of course. If I wrote a novel in the Matrix, and you read it, and so did 10,000 other minds, and I then win the Pulitzer Prize for it, in what sense would it be unreal or even diminished in value? This differs again from the experience machine. In the experience machine, I might have programmed it so that it would *seem* to me that I had written a brilliant novel and that people had appreciated it. In fact, however, no one would have read my novel and I would have simply programmed myself with memories of having written it, although I never really did. In the Matrix, however, I am not given false memories, and I do really interact with other minds. Physics as we know it would be false (not of course the physics of the Matrix, which scientists would study and which would progress as does ordinary science; see below). But art and human relationships would not be affected. I am trying to show that while we are

attached to *reality*, we are not attached to the physical *per se*, where that refers to what we think of as the underlying causes of the smells, tastes, feels, sights, and sounds around us: they could be molecules, they could be computer chips, they could be the whims of Homeric gods. Indeed, very few human beings have much understanding of contemporary physics and what it maintains things "really" are.[14] Nozick's experience machine may have shown us that we have an attachment to the real, an attachment to the truth that we are *really* doing things, *really* accomplishing things, and not just *seeming* to, but we should not for that reason think we are necessarily attached to a certain picture of the physical or metaphysical constitution of things.

I would like return to the question of the sense in which the reality of the Matrix is different from the real world. I think that there is an important difference between being deceived about the reality of an object and being deceived about the real underlying physical or metaphysical cause of something. Avoiding deception and error about the latter is the concern of physics (and metaphysics). That we might be wrong, indeed radically wrong, about the physics/biology of an elephant is quite different from hallucinating that there is an elephant in front of you, or dreaming of an elephant, or experiencing an elephant in Nozick's machine. In the latter three cases, one is deceived about the reality of an object, about whether there is an elephant there at all. I am not saying that the actual physics or metaphysics of a thing will not determine *whether* it is there; if something is really the underlying cause of something else, of course it must determine its existence. I do claim, however, that *given* the reality of a thing, knowing its true physical/metaphysical explanation neither augments nor diminishes its value or its reality.[15] To discover that, contrary to what you had believed, elephants

evolved from single-celled sea creatures and are mostly water, and that water consists of molecules, and that molecules consist of atoms, and that there is a certain interrelationship between matter and energy – that is all part of science's attempt to understand the truth about physical reality. None of these conclusions impugns the elephant's reality or the value it has in the world. What substances at bottom *are* is a question for science or, perhaps, metaphysics. The moral background of the film is quite relevant here. If the fact that we are in the Matrix is simply a matter of our being incorrect about or ignorant of what the real physics of things is, then the Matrix is quite close to our ordinary situation, although our position as *viewers* of *The Matrix* is not like that at all. Since we have a "God's eye" perspective, we are able to know what is really the cause of things and what is not.

In the benevolent Matrix that I envisaged, however, you could learn Matrix-physics and Matrix-history just as we now learn ordinary physics and ordinary history. At a certain age in school you might be taught that your body is really floating in a vat, and then perhaps you could put on goggles and see the world outside of the Matrix, like looking at an x-ray or at your blood under a microscope. Brought up with such a physics and biology, it would seem natural – about as exciting (and unexciting) as being told that your solid unmoving table is made of incredibly small incredibly fast moving parts, or that all of your physical characteristics are determined by a certain code in your DNA, or where babies come from – despite the fact that such truths are hardly obvious, and conflict radically with the way things appear. Just consider any of the conclusions of contemporary physics or quantum mechanics. History too might continue as normal, divided into BM (before Matrix) and AM (after Matrix) dates. After all in the "real" world, outside of the Matrix, nothing would be happening of interest except to scientists. It would be like the contemporary

study of bottom of the ocean, or of the moon. Aside from its causal influence on the physical state of the planet, what goes on down there or up there has no part to play in human history. All of *human* history would occur within the Matrix.

By hypothesizing a benevolent rather than a malevolent cause of the Matrix, we can see how much of what I am calling the "moral background" of *The Matrix* influences what we think of it. Deprived of that moral background, a benevolently generated Matrix can show us that our attachment is not to the physical constitution and cause of things, but also not simply to experience. Our attachment is to things that have *value*. Let me explain.

Take the example, discussed in the film, of the pleasure of eating. Imagine that science develops a pill which supplies the perfect amount of nutrition for a human being each day. Humans no longer need to eat at all in the ordinary way. In fact they are, as far as their health is concerned, far worse off if they try to rely on their taste to supply them with the appropriate nutrition (see current statistics on fast food consumption and obesity). They can simply take the pill and get nutrition far superior to what they would if left to their own taste to determine what and how much to eat. Let's suppose too that science has found a way to simulate food with a computer, so that they have created a "food-matrix". My real nutrition would come from the pill, but I could still go out for a "simulated" steak and it would seem just as though I were really eating a steak, including the sensation of getting full, although in fact I would be eating nothing and getting no nutritional harm or benefit from the experience at all. It is hard to imagine such a perfect pill and such perfect computer-simulated food; such a pill is no simple vitamin, and a tofu-burger is no simulated steak. But if we suppose that there are such things, I think

human beings would readily give up eating real steak. What those who value eating steak value is not the eating of real cow flesh (in fact, putting it that way inclines one to become a vegetarian), but the experience of eating. If eating the computer steak really were, as we are assuming, absolutely indistinguishable from eating a real steak, no one would care whether they were eating a "real" steak – that is, one that was obtained from a slaughtered cow.

At this level the discussion is again about what the underlying causes of phenomenal qualities are: whether the causes of the taste, smell, etc. of the steak are cow molecules or computer chips or the hand of God. This is, as it were, a matter of science or metaphysics – not of concern to the consumer as a consumer. Now for all physical objects, I contend, it is of no value to us if their underlying constitution is ordinary atoms, or computer generated simulation. My favorite pen still writes the same way, my favorite shirt still feels the same way. If these things are not "real" in the sense that their underlying constitution is radically other than I had believed, that makes no difference to the value that these things have in our lives. It *does*, of course, make a difference to the truth of the physics or metaphysics I learn. But none of this implies that I was being deceived about the reality of the object – that the object I valued was or is not there in the sense that matters to the non-scientist.[16] In a scene discussed by Grau, Cypher claims his knowledge that the steak is "unreal" – that is, computer generated – does not diminish his enjoyment. Cypher then looks forward to the point when he expects his memory to be wiped clean, and when he will no longer remember that the Matrix is the Matrix. But it seems to me to be unclear why Cypher needs to forget anything about his steak being unreal in order to fully enjoy it – as he

himself seems to understand – nor does he need to forget that he is in the Matrix in order to make his life pleasant and satisfying within it. What he desperately needs to forget in order to have a comfortable and satisfying life is the memory of his immoral and cowardly betrayal of his friends and of the rest of those outside of the Matrix who are engaged in the fight for human liberation. But this is an issue, once again, not arising from the Matrix itself, but from the "moral background" of the film.

Having a radically different underlying constitution is very different from saying that things are not real, in the sense of being a mere illusion, as in a dream or a hallucination. Consider again the case of our human interactions. If a person I am friends with is not, after all, a person, then I think there is a clear sense in which the friendship is not real, just as in Nozick's experience machine or in a dream that I was friends with Tom Waits. I would then *seem* to have a relationship to someone, but in reality not have one. What matters is whether I am really interacting with another free mind. I certainly won't try to say what it is to have a mind, or what it is for that mind to be "free", but whatever it is, I am claiming that its value is not importantly tied to any theory in physics or metaphysics. *Whatever* the cause and explanation is of the existence of a free mind, it is the having of one and the ability to interact with other ones that matters. If the underlying constitution of Tom Waits is computer chips, instead of blood and guts, what difference does that make? This is not a question about his *reality* – whether he is really there or not –, it is a question about his physical or metaphysical constitution. If he has a mind, whatever that is, and he has free will, whatever that is, what do I care what physical parts he is – or is not – made of?[17] Indeed, I earnestly hope in the actual world never to see any of those parts or have direct contact with them at all.

3. *The Matrix* on the Matrix

I shall conclude by claiming that *The Matrix* itself provides evidence that, barring enslavement and deception, we would prefer life within the Matrix. I have so far considered how we would feel about the reality of a benevolently generated Matrix. But in *The Matrix*, the cause of the Matrix is explicitly not benevolent. Human beings are enslaved and exploited by scary-looking machines. *The Matrix* is a story about a few human beings fighting to save the rest of humanity. That is how the movie generates excitement, the thrill for the viewer as he or she hopes that the heroes can defeat the enemy. Of course, the film expects one to root for the humans. But I think there is some duplicity at work in the way *The Matrix* exploits the Matrix. Neo is the savior of humanity, and a large amount of the pleasure that the viewer gets from the film consists of watching Neo and his friends learn to manipulate the Matrix. Key to Neo's eventual success is his training. In his training he learns that the Matrix, as a computer-generated group dream, can be manipulated by a human being. The idea, I guess, is that if one could bring oneself to believe deeply enough that, despite appearances, things are not real, then one could manipulate the reality of the Matrix. The thrill that Neo feels, and that we feel watching him, is that as he gains this control he is able to do things that are, apparently, superhuman – move faster than bullets, hang onto helicopters, fly, etc. We ought to note here, though, that Neo's greatness, his being the One, is only the case because the Matrix exists. Outside of the Matrix, Neo is just a smart computer geek. He can't really fly, or really dodge bullets (nor, apparently, does he dress in full-length black leather coats, though I guess he *could*). We, as viewers, would not get any pleasure from *The Matrix* if it were not for the Matrix. If there were no Matrix, everyone would be eating terrible

porridge in a sunless world and simply fighting for survival, which would make for a bad world and a bad movie. The premise of the movie is that there is a moral duty to destroy the Matrix, and "free" the humans. But all of the satisfaction that the viewer gets, and that the characters get in terms of their own sense of purpose and of being special, is derived *from* the Matrix. It's not just Cypher's steak that is owed to the Matrix, it is Morpheus's breaking the handcuffs, Trinity's gravity-defying leaps, and Neo's bullet dodging. If my argument is right, then, the irony of *The Matrix* is that the heroes spend all of their time liberating human beings from the Matrix although afterwards they would have good reason to go back in, assuming the conditions on Earth are still so terrible. This is because there's nothing *wrong* with the Matrix *per se*; indeed, I've argued that our reality might just as well *be* the Matrix. What we want, now as always, one way or another, is to have control over it ourselves. What we would do with such power is a question, I suppose, for psychologists; but, looking at what people have done so far, I at any rate hope we remain enslaved and deceived by *something* for a long time to come.[18]

Iakovos Vasiliou

Footnotes

1. Another topic raised by the film, which I will not discuss beyond this note, would be to assess the moral background of the plot. Are the humans clearly in the right? After all, it was they who blotted out the sun in an attempt to exterminate the machines. Particularly in light of the machines' claim that they are simply the next evolutionary step, we ought to think about whether there is some objectionable "speciesism" at work in the humans' assessment of the situation. For my purposes I'll assume the humans are morally justified in the fight for liberation, which, I might add, is certainly a defensible position. For even if machines are the next evolutionary step, and some human beings are guilty of having acted wrongly towards them, that would hardly justify the involuntary enslavement of the entire human race in perpetuity. Moreover, the existence of a "more advanced" species than our own (however that is to be determined) surely should not deprive us of our human rights.

2. And in part because we too would like to control reality; see below.

3. The Stoics thought of the natural world, of the universe as a whole, as itself a rational creature with an overall goal or purpose.

4. Although this theme is present throughout, see especially *Meditations I* and *IV*.

5. The idea that reality is tricky and tries to hide its nature from us is very old, even without, as in Homer's case, any gods acting as agents of deception. For example, the Presocratic philosopher Heraclitus (c.540-c.480 BC) writes (fr. 53) "an unapparent connection is stronger than an apparent one" and (fr. 123) "nature/the real constitution [of things] (*phusis*) loves to hide itself."

6. Morpheus and company are an interesting amalgam of technological sophistication and religious symbolism.

7. Or, more precisely, those of us who accomplished this.

8. Of course for Morpheus and his crew, and for the machines if they were sufficiently reflective, the same questions could be raised about what makes the reality *outside* of the Matrix the way *it* is – who is responsible for *that?* And then we can imagine them responding in the sorts of ways I have described, pinning the blame on God, the laws of nature, etc.

9. This detail is meant simply to avoid the possibility of unease over the issue of whether human beings are being used as batteries, voluntarily or not.

10. I shall assume that my reader has read that essay, where Grau clearly explains Nozick's example. (The essay can be found here.)

11. See Grau, "Dream Skepticism". The threat of solipsism seems to me to be the same in the Matrix or in the ordinary world; and that is not my concern here. I am simply taking the truth of the "God's eye" perspective offered the viewer of the film for granted. *The Matrix* tells us and shows us that we are all hooked up to the same Matrix.

12. I think that perhaps Colin McGinn's essay too quickly assimilates the Matrix to dreaming, and Neo's control over it to "lucid" dreaming. Although McGinn may be right that the Matrix must be dealing with "images" rather than "percepts", there are important disanalogies between Matrix-experience and dream-experience. First, in a dream, there is only your own mind involved. The Matrix must be, at a minimum, a group dream. I am arguing above that the fact that one mind is really interacting with other minds is critical to assessing the value of the Matrix reality. This complicates the apparently clear idea of controlling one's dream, since it is not simply one mind at work that can "alter" the images one is conscious of. I am not sure of the coherence of the hypothesis here. For example, when the young boy bends the spoon, Neo "sees" this. So the boy's control of his environment is perceivable *both* by the boy's mind *and* by Neo's. So he must be changing something that is, "in reality", in Neo's mind – namely, Neo's image of the spoon. But what if Neo straightens the spoon at the same time the boy bends it? Whose lucid dream will win out, and be perceived by the other minds? The one with the stronger will? Second, the "images" that are in your mind in the Matrix can, and regularly do, *really* kill people; that is, kill their bodies *outside* of the Matrix. Except in some bad horror movies, dream images cannot really kill you, or make you bleed. The difficulty of understanding how something which is a mere "image" is supposed to have this sort of effect seems therefore to cause some problems for calling the state of ordinary people in the Matrix "dreaming". See also next note.

13. We could certainly, if we wish, *call* the experience of the Matrix "a dream", as the movie does. But we should remember that Neo, while in the Matrix and before he has met Morpheus, has a dream while he is "asleep". So we need some distinction between that sort of "dream"

and Neo's "waking" "group-dream" within the Matrix.

14. This sentence implies that contemporary physics represents humans' best understanding of the true nature of reality, which is certainly a contentious claim.

15. The question of whether I know something is in fact real or an illusion remains as legitimate or illegitimate as always. As throughout this essay, I am simply bypassing any sceptical questions, since it is part of my argument that being in the Matrix does not affect them.

16. All human being might be considered "scientists" insofar as we are curious about and have a conception of what the reality of things are: what causes them, how they come into being, how they are destroyed, etc. But we are also interested in other people, objects, and activities because of their inherent value, a value they retain regardless of the correct explanation of their reality.

17. Given a true account of what it is to have a mind, I would surely care if what appeared to be a person did not fulfill those criteria, for then he would not be a person after all. For example, if someone somehow showed that a machine could not have a "free mind", then I would care whether my friend was a machine or not, but only secondarily, given that *ex hypothesi* as a machine he would not have a free mind. My point is only that it is "having a free mind" or "being a person" that is the source of value, not the correct theory about what makes someone a person. I am claiming that ignorance of or deception about the right physical or metaphysical account of mind does not thereby cast doubt on the value of having a mind. Scepticism about other minds – the questions of whether there really are other minds and how we could tell whether there are – is not addressed at all by what I am saying. I am taking for granted the truth of what the film tells us: there are other minds. The problem of other minds, like solipsism mentioned above, is equally a problem in or out of the Matrix.

18. I am grateful to Chris Grau and Bill Vasiliou for comments on and discussion about an earlier version of this essay.

The Matrix – Our Future?

Kevin Warwick

Is *The Matrix* merely a science fiction scenario, or is it, rather, a philosophical exercise? Alternatively, is it a realistic possible future world? The number of respected scientists predicting the advent of intelligent machines is growing exponentially. Steven Hawking, perhaps the most highly regarded theoretical scientist in the world and the holder of the Cambridge University chair that once belonged to Isaac Newton, said recently, "In contrast with our intellect, computers double their performance every 18 months. So the danger is real that they could develop intelligence and take over the world." He added, "We must develop as quickly as possible technologies that make possible a direct connection between brain and computer, so that artificial brains contribute to human intelligence rather than opposing it."[1] The important message to take from this is that the danger—that we will see machines with an intellect that outperforms that of humans—is real.

I. The Facts

But is it just a danger—a potential threat—or, if things continue to progress as they are doing, is it an inevitability? Is the Matrix going to happen whether we like it or not? One flaw in the present-day thinking of some philosophers lies in their assumption that the ultimate goal of research into Artificial Intelligence is to create a robot machine with intellectual capabilities approaching those of a human. This may be the aim in a limited number of cases, but the goal for most AI developers is to make use of the ways in which robots can outperform humans—rather than those in which they can only potentially become our

match.

Robots can sense the world in ways that humans cannot—ultraviolet, X-ray, infrared, and ultrasonic perception are some obvious examples—and they can intellectually outperform humans in many aspects of memory and logical mathematical processing. And robots have no trouble thinking of the world around them in multiple dimensions, whereas human brains are still restricted to conceiving the same entity in an extremely limited three dimensional way. But perhaps the biggest advantage robots have over us is their means of communication—generally an electronic form, as opposed to the human's embarrassingly slow mechanical technique called speech, with its highly restricted coding schemes called languages.

It appears to be inevitable that at some stage a sentient robot will appear, its production having been initiated by humans, and begin to produce other, even more capable and powerful robots. One thing overlooked by many is that humans do not reproduce, other than in cloning; rather, humans *produce* other humans. Robots are far superior at producing other robots and can spawn robots that are far more intelligent than themselves.

Once a race of intellectually superior robots has been set into action, major problems will appear for humans. The morals, ethics, and values of these robots will almost surely be drastically different from those of humans. How would humans be able to reason or bargain with such robots? Why indeed should such robots want to take any notice at all of the silly little noises humans would be making? It would be rather like humans today obeying the instructions of cows.

So a war of some kind would be inevitable, in the form of a last gasp from

humans. Even having created intelligent, sentient robots in the first place, robots that can out-think them, the humans' last hope would be to find a weak spot in the robot armoury, a chink in their life-support mechanism. Naturally, their food source would be an ideal target. For the machines, obtaining energy from the sun—a constant source—would let them bypass humans, excluding them from the loop. But as we know, humans have already had much success in polluting the atmosphere and wrecking the ozone layer, so blocking out the sun's rays – scorching the sky, in effect – would seem to be a perfectly natural line of attack in an attempt to deprive machines of energy.

In my own book, *In the Mind of the Machine*[2], I had put forth the idea that the machines would, perhaps in retaliation, use humans as slave labourers, to supply robots with their necessary energy. Indeed, we must consider this as one possible scenario. However, actually using humans as a source of energy—batteries, if you like—is a much sweeter solution, and more complete. Humans could be made to lie in individual pod-like wombs, acting rather like a collection of battery cells, to feed the machine-led world with power.

Probably in this world of machine dominance there would be a few renegade humans causing trouble, snapping at the heels of the machine authorities in an attempt to wrestle back power for humans, an attempt to go back to the good old times. So it is with the Matrix. It is a strange dichotomy of human existence that as a species we are driven by progress—it is central to our being—yet at the same time, for many there is a fruitless desire to step back into a world gone by, a dream world.

Yet it is in human dreams that the Matrix machines have brought about a happy balance. Simply treating humans as slaves would always bring about problems of resistance. But by providing a port directly into each human brain,

each individual can be fed a reality with which he or she is happy, creating for each one a contented existence in a sort of dream world. Even now we know that scientifically it would be quite possible to measure, in a variety of ways, the level of contentment experienced by each person. The only technical problem is how one would go about feeding a storyline directly into a brain.

So what about the practical realities of the brain port? I myself have, as reported in 'I, Cyborg,'[3] had a 100-pin port that allowed for both signal input and output connected into my central nervous system. In one experiment conducted while I was in New York City, signals from my brain, transmitted via the Internet, operated a robot hand in the UK. Meanwhile, signals transmitted onto my nervous system were clearly recognisable in my brain. A brain port, along the lines of that in the Matrix, is not only a scientific best guess for the future; I am working on such a port now, and it will be with us within a decade at most.

II. Human or Machine

With the port connected into my nervous system, my brain was directly connected to a computer and thence on to the network. I considered myself to be a Cyborg: part human, part machine. In *The Matrix*, the story revolves around the battle between humans and intelligent robots. Yet Neo, and most of the other humans, each have their own brain port. When out of the Matrix, they are undoubtedly human; but while they are in the Matrix, there can be no question that they are no longer human, but rather are Cyborgs. The real battle then becomes not one of humans versus intelligent robots but of Cyborgs versus intelligent robots.

The status of an individual whilst within the Matrix raises several key issues. For example, when they are connected are Neo, Morpheus, and Trinity individuals within the Matrix? Or do they have brains which are part human, part machine? Are they themselves effectively a node on the Matrix, sharing common brain elements with others? It must be remembered that ordinarily human brains operate in a stand-alone mode, whereas computer-brained robots are invariably networked. When connected into a network, as in the Matrix, and as in my own case as a Cyborg, individuality takes on a different form. There is a unique, usually human element, and then a common, networked machine element.

Using the common element, 'reality' can be downloaded into each brain. Morpheus describes this (as do others throughout the film) as 'having a dream.' He raises questions as to what is real. He asks how it is possible to know the difference between the dream world and the real world. This line of questioning follows on from many philosophical discussions, perhaps the most prominent being that of Descartes, who appeared to want to make distinctions between dream states and 'reality', immediately leading to problems in defining what was real and what was not. As a result he faced further problems in defining absolute truths.

Perhaps a more pertinent approach can be drawn from Berkeley, who denied the existence of a physical world, and Nietzsche, who scorned the idea of objective truth. By making the basic assumption that there is no God, my own conclusion is that *there can be no absolute reality, there can be no absolute truth* — whether we be human, Cyborg, or robot. Each individual brain draws its conclusions and makes assumptions as to the reality it faces at an instant, dependant on the input it receives. If only limited sensory input is forthcoming, then brain memory banks (or injected feelings) need to be tapped for a brain

to conceive of a storyline. At any instant, a brain links its state with its common-sense memory banks, often coming to unlikely conclusions.

As a brain ages, or as a result of an accident, the brain's workings can change; this often appears to the individual to be a change in what is perceived rather than a change in that which is perceiving. In other words, the individual thinks it must be the world that has changed, not his or her brain. Where a brain is part of a network, however, there is a possibility for alternative viewpoints to be proposed by different nodes on the network. This is not something that individual humans are used to. An individual brain tends to draw only one conclusion at a time. In some types of schizophrenia this conclusion can be confused and can change over time; it is more usually the case, though, that such an individual will draw a conclusion about what is perceived that is very much at variance with the conclusion of other individuals. For the most part, what is deemed by society to be 'reality' at any point, far from being an absolute, is merely a commonly agreed set of values based on the perceptions of a group of individuals.

The temptation to see a religious undertone in *The Matrix* is interesting — with Morpheus cast as the prophet John the Baptist, Trinity perhaps as God or the holy spirit, Neo clearly as the messiah, and Cypher as Judas Iscariot, the traitor. But, far from a Gandhi-like, turn the other cheek, approach, Neo's is closer to one that perhaps was actually expected by many of the messiah himself, taking on his role as victor over the evil Matrix: a holy war against a seemingly invincible, all-powerful machine network.

But what of the machine network, the Matrix, itself? With an intellect well above that of collective humanity, surely its creativity, its artistic sense, its

value for aesthetics would be a treat to behold. But the film keeps this aspect from us – perhaps to be revealed in a sequel. Humans released from the Matrix grip, merely regard it as an evil, perhaps Cypher excluded here. Meanwhile the Agents are seen almost as faceless automatons, ruthless killers, strictly obeying the will of their Matrix overlord. Possibly humans would see both the Matrix and Agents as the enemy, just as the Matrix and Agents would so regard humans – but once inside the Matrix the picture is not so clear. As a Cyborg, who are your friends and who are your enemies? It is no longer black and white when you are part machine, part human.

III. In and Out of Control

Morpheus tells Neo that the Matrix is control. This in itself is an important revelation. As humans, we are used to one powerful individual being the main instigator, the brains behind everything. It is almost as though we cannot even conceive of a group or collection running amuck, but believe, rather, that there is an individual behind it all. In the second world war, it was not the Germans or Germany who the allies were fighting but Adolf Hitler; meanwhile in Afghanistan, it is Bin-Laden who is behind it all. Yet in the Matrix we are faced with a much more realistic scenario, in that it is not some crazed individual up to no good, but the Matrix – a network.

When I find myself in a discussion of the possibility of intelligent machines taking over things, nine times out of ten I am told—following a little chuckle to signify that I have overlooked a blindingly obvious point—that "If a machine causes a problem you can always switch it off." What a fool I was not to have thought of it!! How could I have missed that little snippet?

Of course it is not only the Matrix but even today’s common Internet that gives

us the answer, and cuts the chuckle short. Even now, how is it practically possible to switch off the Internet? We're not talking theory here, we're talking practice. Okay, it is of course possible to unplug one computer, or even a small subsection intranet, but to bring down the whole Internet? Of course we can't. Too many entities, both humans and machines, rely on its operation for their everyday existence. It is not a Matrix of the future that we will not be able to switch off, it is a Matrix of today that we cannot switch off, over which we cannot have ultimate control.

Neo learns that the Matrix is a computer-generated dream world aimed at keeping humans under control. Humans are happy to act as an energy source for the Matrix as long as they themselves believe that the reality of their existence is to their liking; indeed, how are the human nodes in a position to know what is computer-generated reality and what is reality generated in some other way?

A stand-alone human brain operates electrochemically, powered partly by electrical signals and partly by chemicals. In the western world we are more used to chemicals being used to change our brain and body state, either for medicinal purposes or through narcotics, including chemically instigated hallucinations. But now we are entering the world of e-medicine. Utilising the electronic element of the electrochemical signals on which the human brain and nervous system operate, counterbalancing signals can be sent to key nerve fibre groups to overcome a medical problem. Conversely, electronics signals can be injected to stimulate movement or pleasure. Ultimately, electronic signals will be able to replace the chemicals that release memories and "download" memories not previously held. Why live in a world that is not to your liking if a Matrix state is able to keep your bodily functions operating

whilst you live out a life in a world in which you are happy with yourself? The world of the Matrix would appear to be one that lies in the direction humanity is now heading—a direction in which it would seem, as we defer more and more to machines to make up our minds for us, that we wish to head.

IV. Ignorance and Bliss

In a sense, The Matrix is nothing more than a modern day "Big Brother," taking on a machine form rather than the Orwellian vision of a powerful individual using machines to assist and bring about an all-powerful status. But *1984*, the novel in which the story of Big Brother was presented, was published in 1948. The Matrix comes fifty years later. In the meantime, we have witnessed the likes of radar, television for all, space travel, computers, mobile phones, and the Internet. What would Orwell's Big Brother have been like if he had had those technologies at his disposal – would Big Brother have been far from the Matrix?

With the first implant I received, in 1998, for which I had no medical reason (merely scientific curiosity), a computer network was able to monitor my movements. It knew what time I entered a room and when I left. In return it opened doors for me, switched on lights, and even gave me a welcoming "Hello" as I arrived. I experienced no negatives at all. In fact, I felt very positive about the whole thing. I gained something as a result of being monitored and tracked. I was happy with having Big Brother watching me because, although I gave up some of my individual humanity, I benefited from the system doing things for me. Would the same not be true of the Matrix? Why would anyone want to experience the relatively tough and dangerous life of being an individual human when he or she could be part of the Matrix?

So here we come on to the case of Cypher. As he eats his steak he says, "I know that this steak doesn't exist. I know when I put it in my mouth, the Matrix is telling my brain that it is juicy and delicious!" He goes on to conclude that "Ignorance is bliss." But is it ignorance? His brain is telling him, by whatever means, that he is eating a nice juicy steak. How many times do we nowadays enter a fast-food burger bar in order to partake of a burger that, through advertising, our brains have been conditioned into believing is the tastiest burger imaginable. When we enter we know, because we've seen the scientific papers, that the burger contains a high percentage of water, is mainly fat, and is devoid of vitamins. Yet we still buy such burgers by the billion. When we eat one, our conditioned brain is somehow telling us that it is juicy and delicious, yet we know it doesn't quite exist in the form our brain is imagining.

We can thus understand Cypher's choice. Why be out of the Matrix, living the dangerous, poor, tired, starving life of a disenfranchised human, when you can exist in a blissfully happy life, with all the nourishment you need? Due to the deal he made with Agent Smith, once Cypher is back inside he will have no knowledge of having made any deal in the first place. He appears to have nothing at all to lose. The only negative aspect is that before he is reinserted he may experience some inner moral human pangs of good or bad. Remember that being reinserted is actually good for the Matrix, although it is not so good for the renegade humans who are fighting the system.

Robert Nozick's thought experiment puts us all to the test, and serves as an immediate exhibition of Cypher's dilemma. Nozick asks, if our brains can be connected, by electrodes, to a machine which gives us any experiences we desire, would we plug into it for life? The question is, what else could matter

other than how we feel our lives are going, from the inside? Nozick himself argued that other things do matter to us, for example that we value being a certain type of person, we want to be decent, we actually wish to do certain things rather than just have the experience of doing them. I disagree completely with Nozick.

Research involving a variety of creatures, principally chimpanzees and rats, has allowed them to directly stimulate pleasure zones in their own brain, simply by pressing a button. When given the choice of pushing a button for pleasure or a button for food, it is the pleasure button that has been pressed over and over again, even leading to starvation (although individuals were quite happy even about that). Importantly, the individual creatures still had a role to play, albeit merely that of pressing a button. This ties in directly with the Matrix, which also allows for each individual mentally experiencing a world in which he or she is active and has a role to play.

It is, however, an important question whether or not an individual, as part of the Matrix, experiences free will or not. It could be said that Cypher, in deciding to re-enter the Matrix, is exercising his free will. But once inside, will he still be able to exhibit free will then? Isn't it essentially a similar situation to that proposed by Nozick? Certainly, within the mental reality projected on an individual by the Matrix, it is assumed that a certain amount of mental free will is allowed for; but it must be remembered, at the same time, that each individual is lying in a pod with all his or her life-sustaining mechanisms taken care of and an interactive storyline being played down into his or her brain. Is that free will? What is free will anyway, when the state of a human brain is merely partly due to a genetic program and partly due to life's experience? Indeed, exactly the same thing is true for a robot.

In the Matrix, no human fuel cells are killed, not even the unborn—there is no abortion. Yet, naturally dying humans are allowed to die naturally and are used as food for the living. Importantly, they are not kept alive by chemicals merely for the sake of keeping them alive. The Matrix would appear to be more morally responsible to its human subjects than are human subjects to themselves. Who therefore wouldn't want to support and belong to the Matrix, especially when it is making life easier for its subjects?

Neo is kidnapped by Luddites, dinosaurs from the past when humans ruled the earth. It's not the future. We are in reality heading towards a world run by machines with an intelligence far superior to that of an individual human. But by linking into the network and becoming a Cyborg, life can appear to be even better than it is now. We really need to clamp down on the party-pooper Neos of this world and get into the future as soon as we can—a future in which we can be part of a Matrix system, which is morally far superior to our Neolithic morals of today.

Kevin Warwick

www.kevinwarwick.com

Footnotes

1. Hawking, S., "Hawking's plan to offest computer threat to humans", Ananova, www.ananova/news, 1 September 2001

2. *In the Mind of the Machine*, Arrow, 1998. Available on www.amazon.co.uk

3. *I, Cyborg*, Century, 2002. Available on www.amazon.co.uk

Wake Up!
Gnosticism & Buddhism in *The Matrix*

Frances Flannery-Dailey & Rachel Wagner

At the beginning of *The Matrix*, a black-clad computer hacker known as Neo falls asleep in front of his computer. A mysterious message appears on the screen: "Wake up, Neo."[1] This succinct phrase encapsulates the plot of the film, as Neo struggles with the problem of being imprisoned in a "material" world that is actually a computer simulation program created in the distant future by Artificial Intelligence (AI) as a means of enslaving humanity, by perpetuating ignorance in the form of an illusory perception called "the Matrix." In part, the film crafts its ultimate view of reality by alluding to numerous religious traditions that advance the idea that the fundamental problem which humanity faces is ignorance and the solution is knowledge or awakening. Two religious traditions on which the film draws heavily are Gnostic Christianity and Buddhism.[2] Although these traditions differ in important ways, they agree in maintaining that the problem of ignorance can be solved through an individual's reorientation of perspective concerning the material realm.[3] Gnostic Christianity and Buddhism also both envision a guide who helps those still trapped in the limiting world of illusion, a Gnostic redeemer figure or a bodhisattva, who willingly enters that world in order to share liberating knowledge, facilitating escape for anyone able to understand. In the film, this figure is Neo, whose name is also an anagram for the "One."

Although as a "modern myth"[4] the film purposefully draws on numerous traditions,[5] we propose that an examination of Gnostic Christianity and Buddhism well illuminates the overarching paradigm of *The Matrix*, namely, the

problem of sleeping in ignorance in a dreamworld, solved by waking to knowledge or enlightenment. By drawing syncretistically on these two ancient traditions and fusing them with a technological vision of the future, the film constructs a new teaching that challenges its audience to question "reality."

I. Christian Elements in *The Matrix*

The majority of the film's audience probably easily recognizes the presence of some Christian elements, such as the name Trinity[6] or Neo's death and Christ-like resurrection and ascension near the end of the film. In fact, Christian and biblical allusions abound, particularly with respect to nomenclature:[7] Apoc (Apocalypse), Neo's given name of Mr. Ander/son (from the Greek andras for man, thus producing "Son of Man"), the ship named the Nebuchadnezzar (the Babylonian king who, in the Book of Daniel, has puzzling symbolic dreams that must be interpreted),[8] and the last remaining human city, Zion, synonymous in Judaism and Christianity with (the heavenly) Jerusalem.[9] Neo is overtly constructed as a Jesus figure: he is "the One" who was prophesied to return again to the Matrix, who has the power the change the Matrix from within (i.e., to work miracles), who battles the representatives of evil and who is killed but comes to life again.

This construction of Neo as Jesus is reinforced in numerous ways. Within minutes of the commencement of the movie, another hacker says to Neo, "You're my savior, man, my own personal Jesus Christ."[10] This identification is also suggested by the Nebuchadnezzar's crew, who nervously wonder if he is "the One" who was foretold, and who repeatedly swear in Neo's presence by saying "Jesus" or "Jesus Christ."[11] In still another example, Neo enters the Nebuchadnezzar for the first time and the camera pans across the interior of

the ship, resting on the make: "Mark III no. 11." This seems to be another messianic reference, since the Gospel of Mark 3:11 reads: "Whenever the unclean spirits saw him, they fell down before him and shouted, ' You are the Son of God!'"

Like Mark's Jesus, Neo is an exorcist, who casts out alien Agents inhabiting the residual self-images of those immersed in the Matrix. However, this trope illuminates the differences between Jesus and Neo, since the latter accomplishes exorcisms not by healing, but by killing the digital bodies of those who are "possessed" by Agents, in turn killing the real people in the world of the Nebuchadnezzar. The plaque, then, ultimately highlights the problem of violence in the film, even as it draws parallels between Jesus and Neo.

II. Gnosticism in *The Matrix*

Although the presence of individual Christian elements within the film is clear, the overall system of Christianity that is presented is not the traditional, orthodox one. Rather, the Christian elements of the film make the most sense when viewed within a context of Gnostic Christianity.[12] Gnosticism was a religious system that flourished for centuries at the beginning of the Common Era, and in many regions of the ancient Mediterranean world it competed strongly with "orthodox" Christianity, while in other areas it represented the only interpretation of Christianity that was known.[13] The Gnostics possessed their own Scriptures, accessible to us in the form of the Nag Hammadi Library, from which a general sketch of Gnostic beliefs may be drawn.[14] Although Gnostic Christianity comprises many varieties, Gnosticism as a whole seems to have embraced an orienting cosmogonic myth that explains the true nature of the universe and humankind's proper place in it.[15] A brief retelling of this myth

illuminates numerous parallels with The Matrix.

In the Gnostic myth, the supreme god is completely perfect and therefore alien and mysterious, "ineffable," "unnamable," "immeasurable light which is pure, holy and immaculate" (Apocryphon of John). In addition to this god there are other, lesser divine beings in the pleroma (akin to heaven, a division of the universe that is not Earth), who possess some metaphorical gender of male or female.[16] Pairs of these beings are able to produce offspring that are themselves divine emanations, perfect in their own ways.[17] A problem arises when one "aeon" or being named Sophia (Greek for wisdom), a female, decides "to bring forth a likeness out of herself without the consent of the Spirit," that is, to produce an offspring without her consort (Apocry. of John). The ancient view was that females contribute the matter in reproduction, and males the form; thus, Sophia's action produces an offspring that is imperfect or even malformed, and she casts it away from the other divine beings in the pleroma into a separate region of the cosmos. This malformed, ignorant deity, sometimes named Yaldaboath, mistakenly believes himself to be the only god.

Gnostics identify Yaldabaoth as the Creator God of the Old Testament, who himself decides to create archons (angels), the material world (Earth) and human beings. Although traditions vary, Yaldabaoth is usually tricked into breathing the divine spark or spirit of his mother Sophia that formerly resided in him into the human being (especially Apocry. of John; echoes of Genesis 2-3). Therein lies the human dilemma. We are pearls in the mud, a divine spirit (good) trapped in a material body (bad) and a material realm (bad). Heaven is our true home, but we are in exile from the pleroma.

Luckily for the Gnostic, salvation is available in the form of gnosis or knowledge imparted by a Gnostic redeemer, who is Christ, a figure sent from

the higher God to free humankind from the Creator God Yaldabaoth. The gnosis involves an understanding of our true nature and origin, the metaphysical reality hitherto unknown to us, resulting in the Gnostic's escape (at death) from the enslaving material prison of the world and the body, into the upper regions of spirit. However, in order to make this ascent, the Gnostic must pass by the archons, who are jealous of his/her luminousity, spirit or intelligence, and who thus try to hinder the Gnostic's upward journey.

To a significant degree, the basic Gnostic myth parallels the plot of *The Matrix*, with respect to both the problem that humans face as well as the solution. Like Sophia, we conceived an offspring out of our own pride, as Morpheus explains: "Early in the 21st century, all of mankind was united in celebration. We marveled at our own magnificence as we gave birth to AI."[18] This offspring of ours, however, like Yaldabaoth is malformed (matter without spirit?). Morpheus describes AI as "a singular consciousness that spawned an entire race of machines," a fitting parallel for the Gnostic Creator God of the archons (angels) and the illusory material world. AI creates the Matrix, a computer simulation that is "a prison for your mind." Thus, Yaldabaoth/ AI traps humankind in a material prison that does not represent ultimate reality, as Morpheus explains to Neo: "As long as the Matrix exists, the human race will never be free."

The film also echoes the metaphorical language employed by Gnostics. The Nag Hammadi texts describe the fundamental human problem in metaphorical terms of blindness, sleep, ignorance, dreams and darkness / night, while the solution is stated in terms of seeing, waking, knowledge (gnosis), waking from dreams and light / day.[19]

Similarly, in the film Morpheus, whose name is taken from the Greek god of

sleep and dreams, reveals to Neo that the Matrix is "a computer generated dreamworld." When Neo is unplugged and awakens for the first time on the Nebuchadnezzar in a brightly lit white space (a cinematic code for heaven), his eyes hurt, as Morpheus explains, because he has never used them. Everything Neo has "seen" up to that point was seen with the mind's eye, as in a dream, created through software simulation. Like an ancient Gnostic, Morpheus explains that the blows he deals Neo in the martial arts training program have nothing to do with his body or speed or strength, which are illusory. Rather, they depend only on his mind, which is real.

The parallels between Neo and Christ sketched earlier are further illuminated by a Gnostic context, since Neo is "saved" through gnosis or secret knowledge, which he passes on to others. Neo learns about the true structure of reality and about his own true identity, which allows him to break the rules of the material world he now perceives to be an illusion. That is, he learns that "the mind makes it [the Matrix, the material world] real," but it is not ultimately real. In the final scene of the film, it is this gnosis that Neo passes on to others in order to free them from the prison of their minds, the Matrix. He functions as a Gnostic Redeemer, a figure from another realm who enters the material world in order to impart saving knowledge about humankind's true identity and the true structure of reality, thereby setting free anyone able to understand the message.

In fact, Neo's given name is not only Mr. Anderson / the Son of Man, it is Thomas Anderson, which reverberates with the most famous Gnostic gospel, the Gospel of Thomas. Also, before he is actualized as Neo (the one who will initiate something "New," since he is indeed "the One"), he is doubting Thomas, who does not believe in his role as the redeemer figure.[20] In fact, the

name Thomas means "the Twin," and in ancient Christian legend he is Jesus' twin brother. In a sense, the role played by Keanu Reeves has a twin character, since he is constructed as both a doubting Thomas and as a Gnostic Christ figure.[21]

Not only does Neo learn and pass on secret knowledge that saves, in good Gnostic fashion, but the way in which he learns also evokes some elements of Gnosticism. Imbued with images from eastern traditions, the training programs teach Neo the concept of "stillness," of freeing the mind and overcoming fear, cinematically captured in "Bullet Time" (digitally mastered montages of freeze frames / slow motion frames using multiple cameras).[22] Interestingly enough, this concept of "stillness" is also present in Gnosticism, in that the higher aeons are equated with "stillness" and "rest" and can only be apprehended in such a centered and meditative manner, as is apparent in these instructions to a certain Allogenes: "And although it is impossible for you to stand, fear nothing; but if you wish to stand, withdraw to the Existence, and you will find it standing and at rest after the likeness of the One who is truly at rest...And when you becomes perfect in that place, still yourself... " (Allogenes) The Gnostic then reveals, "There was within me a stillness of silence, and I heard the Blessedness whereby I knew my proper self" (Allogenes).[23] When Neo realizes the full extent of his "saving gnosis," that the Matrix is only a dreamworld, a reflective Keanu Reeves silently and calmly contemplates the bullets that he has stopped in mid-air, filmed in "Bullet Time."

Yet another parallel with Gnosticism occurs in the portrayal of the Agents such as Agent Smith, and their opposition to the equivalent of the Gnostics - that is, Neo and anyone else attempting to leave the Matrix. AI created these artificial programs to be "the gatekeepers - they are guarding all the doors, they are

holding all the keys." These Agents are akin to the jealous archons created by Yaldabaoth who block the ascent of the Gnostic as he/she tries to leave the material realm and guard the gates of the successive levels of heaven (e.g., Apocalypse of Paul).[24]

However, as Morpheus predicts, Neo is eventually able to defeat the Agents because while they must adhere to the rules of the Matrix, his human mind allows him to bend or break these rules.[25] Mind, though, is not equated in the film merely with rational intelligence, otherwise Artificial Intelligence would win every time. Rather, the concept of "mind" in the film appears to point to a uniquely human capacity for imagination, for intuition, or, as the phrase goes, for "thinking outside the box." Both the film and the Gnostics assert that the "divine spark" within humans allows a perception of gnosis greater than that achievable by even the chief archon / agent of Yaldabaoth:

> And the power of the mother [Sophia, in our analogy, humankind] went out of Yaldabaoth [AI] into the natural body which they had fashioned [the humans grown on farms by AI]... And in that moment the rest of the powers [archons / Agents] became jealous, because he had come into being through all of them and they had given their power to the man, and his intelligence ["mind"] was greater than that of those who had made him, and greater than that of the chief archon [Agent Smith?]. And when they recognized that he was luminous, and that he could think better than they... they took him and threw him into the lowest region of all matter [simulated by the Matrix]. (Apocry. of John 19-20)

It is striking that Neo overcomes Agent Smith in the final showdown of the film precisely by realizing fully the illusion of the Matrix, something the Agent apparently cannot do, since Neo is subsequently able to break rules that the Agent cannot. His final defeat of Smith entails entering Smith's body and splitting him in pieces by means of pure luminosity, portrayed through special effects as light shattering Smith from the inside out.

Overall, then, the system portrayed in *The Matrix* parallels Gnostic Christianity in numerous respects, especially the delineation of humanity's fundamental problem of existing in a dreamworld that simulates reality and the solution of waking up from illusion. The central mythic figures of Sophia, Yaldabaoth, the archons and the Gnostic Christ redeemer also each find parallels with key figures in the film and function in similar ways. The language of Gnosticism and the film are even similar: dreaming vs. waking; blindness vs. seeing;[26] light vs. dark.[27]

However, given that Gnosticism presumes an entire unseen realm of divine beings, where is God in the film? In other words, when Neo becomes sheer light, is this a symbol for divinity, or for human potential? The question becomes even more pertinent with the identification of humankind with Sophia - a divine being in Gnosticism. On one level, there appears to be no God in the film. Although there are apocalyptic motifs, Conrad Ostwalt rightly argues that unlike conventional Christian apocalypses, in *The Matrix* both the catastrophe and its solution are of human making - that is, the divine is not apparent.[28] However, on another level, the film does open up the possibility of a God through the figure of the Oracle, who dwells inside the Matrix and yet has access to information about the future that even those free from the Matrix do not possess. This suggestion is even stronger in the original screenplay, in which the Oracle's apartment is the Holy of Holies nested within the "Temple of Zion."[29] Divinity may also play a role in Neo's past incarnation and his coming again as the One. If, however, there is some implied divinity in the film,[30] it remains transcendent, like the divinity of the ineffable, invisible supreme god in Gnosticism, except where it is immanent in the form of the divine spark active in humans.[31]

III. Buddhism in *The Matrix*

When asked by a fan if Buddhist ideas influenced them in the production of the movie, the Wachowski brothers offered an unqualified "Yes."[32] Indeed, Buddhist ideas pervade the film and appear in close proximity with the equally strong Christian imagery. Almost immediately after Neo is identified as "my own personal Jesus Christ," this appellation is given a distinctively Buddhist twist. The same hacker says: "This never happened. You don't exist." From the stupa-like[33] pods which encase humans in the horrific mechanistic fields to Cypher's selfish desire for the sensations and pleasures of the Matrix, Buddhist teachings form a foundation for much of the film's plot and imagery.[34]

The Problem of Samsara. Even the title of the film evokes the Buddhist worldview. The Matrix is described by Morpheus as "a prison for your mind." It is a dependent "construct" made up of the interlocking digital projections of billions of human beings who are unaware of the illusory nature of the reality in which they live and are completely dependent on the hardware attached to their real bodies and the elaborate software programs created by AI This "construct" resembles the Buddhist idea of samsara, which teaches that the world in which we live our daily lives is constructed only from the sensory projections formulated from our own desires. When Morpheus takes Neo into the "construct" to teach him about the Matrix, Neo learns that the way in which he had perceived himself in the Matrix was nothing more than "the mental projection of your digital self." The "real" world, which we associate with what we feel, smell, taste, and see, "is simply electrical signals interpreted by your brain." The world, Morpheus explains, exists "now only as part of a neural interactive simulation that we call the Matrix." In Buddhist terms, we could say that "because it is empty of self or of what belongs to self, it is therefore said:

'The world is empty.' And what is empty of self and what belongs to self? The eye, material shapes, visual consciousness, impression on the eye -- all these are empty of self and of what belongs to self."[35] According to Buddhism and according to *The Matrix*, the conviction of reality based upon sensory experience, ignorance, and desire keeps humans locked in illusion until they are able to recognize the false nature of reality and relinquish their mistaken sense of identity.

Drawing upon the Buddhist doctrine of Dependent Co-Origination, the film presents reality within the Matrix as a conglomerate of the illusions of all humans caught within its snare. Similarly, Buddhism teaches that the suffering of human beings is dependent upon a cycle of ignorance and desire which locks humans into a repetitive cycle of birth, death, and rebirth. The principle is stated in a short formula in the Samyutta-nikaya:

> If this is that comes to be;
> from the arising of this that arises;
> if this is not that does not come to be;
> from the stopping of this that is stopped.[36]

The idea of Dependent Co-Origination is illustrated in the context of the film through the illusion of the Matrix. The viability of the Matrix's illusion depends upon the belief by those enmeshed in it that the Matrix itself is reality. AI's software program is, in and of itself, no illusion at all. Only when humans interact with its programs do they become enmeshed in a corporately-created illusion, the Matrix, or samsara, which reinforces itself through the interactions of those beings involved within it. Thus the Matrix's reality only exists when actual human minds subjectively experience its programs.[37]

The problem, then, can be seen in Buddhist terms. Humans are trapped in a

cycle of illusion, and their ignorance of this cycle keeps them locked in it, fully dependent upon their own interactions with the program and the illusions of sensory experience which these provide, and the sensory projections of others. These projections are strengthened by humans' enormous desire to believe that what they perceive to be real is in fact real. This desire is so strong that it overcomes Cypher, who can no longer tolerate the "desert of the real" and asks to be reinserted into the Matrix. As he sits with Agent Smith in an upscale restaurant smoking a cigar with a large glass of brandy, Cypher explains his motives:

> "You know, I know this steak doesn't exist. I know that when I put it in my mouth, the Matrix is telling my brain that it is juicy and delicious. After nine years, you know what I realize? Ignorance is bliss."[38]

Cypher knows that the Matrix is not real and that any pleasures he experiences there are illusory. Yet for him, the "ignorance" of samsara is preferable to enlightenment. Denying the reality that he now experiences beyond the Matrix, he uses the double negative: "I don't want to remember nothing. Nothing. And I want to be rich. Someone important. Like an actor." Not only does Cypher want to forget the "nothing" of true reality, but he also wants to be an "actor," to add another level of illusion to the illusion of the Matrix that he is choosing to re-enter.[39] The draw of samsara is so strong that not only does Cypher give in to his cravings, but Mouse also may be said to have been overwhelmed by the lures of samsara, since his death is at least in part due to distractions brought on by his sexual fantasies about the "woman in the red dress" which occupy him when he is supposed to be standing alert.

Whereas Cypher and Mouse represent what happens when one gives in to samsara, the rest of the crew epitomize the restraint and composure praised

by the Buddha. The scene shifts abruptly from the restaurant to the mess hall of the Nebuchadnezzar, where instead of being offered brandy, cigars and steak, Neo is given the "bowl of snot" which is to be his regular meal from that point forward. In contrast to the pleasures which for Cypher can only be fulfilled in the Matrix, Neo and the crew must be content with the "single-celled protein combined with synthetic aminos, vitamins, and minerals" which Dozer claims is "everything the body needs." Clad in threadbare clothes, subsisting on gruel, and sleeping in bare cells, the crew is depicted enacting the Middle Way taught by the Buddha, allowing neither absolute asceticism nor indulgence to distract them from their work.[40]

The Solution of Knowledge/Enlightenment. This duality between the Matrix and the reality beyond it sets up the ultimate goal of the rebels, which is to free all minds from the Matrix and allow humans to live out their lives in the real world beyond. In making this point, the film-makers draw on both Theravada and Mahayana Buddhist ideas.[41] Alluding to the Theravada ideal of the arhat, the film suggests that enlightenment is achieved through individual effort.[42] As his initial guide, Morpheus makes it clear that Neo cannot depend upon him for enlightenment. Morpheus explains, "No one can be told what the Matrix is. You have to see it for yourself." Morpheus tells Neo he must make the final shift in perception entirely on his own. He says: "I'm trying to free your mind, Neo. But I can only show you the door. You're the one that has to walk through it." For Theravada Buddhists, "man's emancipation depends on his own realization of the Truth, and not on the benevolent grace of a god or any external power as a reward for his obedient good behavior."[43] The Dhammapada urges the one seeking enlightenment to "Free thyself from the past, free thyself from the future, free thyself from the present. Crossing to the farther shore of existence, with mind released everywhere, no more shalt thou

come to birth and decay."[44] As Morpheus says to Neo, "There's a difference between knowing the path and walking the path." And as the Buddha taught his followers, "You yourselves should make the effort; the Awakened Ones are only teachers."[45] As one already on the path to enlightenment, Morpheus is only a guide; ultimately Neo must recognize the truth for himself.

Yet *The Matrix* also embraces ideas found in Mahayana Buddhism, especially in its particular concern for liberation for all people through the guidance of those who remain in samsara and postpone their own final enlightenment in order to help others as bodhisattvas.[46] The crew members of the Nebuchadnezzar epitomize this compassion. Rather than remain outside of the Matrix where they are safer, they choose to re-enter it repeatedly as ambassadors of knowledge with the ultimate goal of freeing the minds and eventually also the bodies of those who are trapped within the Matrix's digital web. The film attempts to blend the Theravada ideal of the arhat with the Mahayana ideal of the bodhisattva, presenting the crew as concerned for those still stuck in the Matrix and willing to re-enter the Matrix to help them, while simultaneously arguing that final realization is an individual process.

Neo as the Buddha. Although the entire crew embodies the ideals of the bodhisattva, the filmmakers set Neo apart as unique, suggesting that while the crew may be looked at as arhats and bodhisattvas, Neo can be seen as a Buddha. Neo's identity as the Buddha is reinforced not only through the anagram of his name but also through the myth that surrounds him. The Oracle has foretold the return of one who has the ability to manipulate the Matrix. As Morpheus explains, the return of this man "would hail the destruction of the Matrix, end the war, bring freedom to our people. That is why there are those of us who have spent our entire lives searching the Matrix,

looking for him." Neo, Morpheus believes, is a reincarnation of that man and like the Buddha, he will be endowed with extraordinary powers to aid in the enlightenment of all humanity.

The idea that Neo can be seen as a reincarnation of the Buddha is reinforced by the prevalence of birth imagery in the film directly related to him. At least four incarnations are perceptible in the film. The first birth took place in the pre-history of the film, in the life and death of the first enlightened one who was able to control the Matrix from within. The second consists of Neo's life as Thomas Anderson. The third begins when Neo emerges, gasping, from the gel of the eerily stupa-like pod in which he has been encased, and is unplugged and dropped through a large black tube which can easily be seen as a birth canal.[47] He emerges at the bottom bald, naked, and confused, with eyes that Morpheus tells him have "never been used" before. Having "died" to the world of the Matrix, Neo has been "reborn" into the world beyond it. Neo's fourth life begins after he dies and is "reborn" again in the closing scenes of the film, as Trinity resuscitates him with a kiss.[48] At this point, Neo perceives not only the limitations of the Matrix, but also the limitations of the world of the Nebuchadnezzar, since he overcomes death in both realms. Like the Buddha, his enlightenment grants him omniscience and he is no longer under the power of the Matrix, nor is he subject to birth, death, and rebirth within AI's mechanical construct.[49]

Neo, like the Buddha, seeks to be free from the Matrix and to teach others how to free themselves from it as well, and any use of superhuman powers are engaged to that end. As the only human being since the first enlightened one who is able to freely manipulate the software of the Matrix from within its confines, Neo represents the actualization of the Buddha-nature, one who can

not only recognize the "origin of pain in the world of living beings," but who can also envision "the stopping of the pain," enacting "that course which leads to its stopping."[50] In this sense, he is more than his bodhisattva companions, and offers the hope of awakening and freedom for all humans from the ignorance that binds them.

The Problem of *Nirvana*. But what happens when the Matrix's version of reality is dissolved? Buddhism teaches that when samsara is transcended, nirvana is attained. The notion of self is completely lost, so that conditional reality fades away, and what remains, if anything, defies the ability of language to describe. In his re-entry into the Matrix, however, Neo retains the "residual self-image" and the "mental projection of [a] digital self." Upon "enlightenment," he finds himself not in nirvana, or no-where, but in a different place with an intact, if somewhat confused, sense of self which strongly resembles his "self" within the Matrix. Trinity may be right that the Matrix "cannot tell you who you are," but who you are seems to be at least in some sense related to who you think you are in the Matrix. In other words, there is enough continuity in self-identity between the world of the Matrix and "the desert of the real" that it seems probable that the authors are implying that full "enlightenment" has not yet been reached and must lie beyond the reality of the Nebuchadnezzar and the world it inhabits. If the Buddhist paradigm is followed to its logical conclusions, then we have to expect at least one more layer of "reality" beyond the world of the crew, since even freed from the Matrix they are still subject to suffering and death and still exhibit individual egos.

This idea is reinforced by what may be the most problematic alteration which *The Matrix* makes to traditional Buddhist teachings. The Buddhist doctrine of

ahimsa, or non-injury to all living beings, is overtly contradicted in the film.[51] It appears as if the filmmakers deliberately chose to link violence with salvific knowledge, since there seems to be no way that the crew could succeed without the help of weaponry. When Tank asks Neo and Trinity what they need for their rescue of Morpheus "besides a miracle," their reply is instantaneous: "Guns -- lots of guns." The writers could easily have presented the "deaths" of the Agents as nothing more than the ending of that particular part of the software program. Instead, the Wachowski brothers have purposefully chosen to portray humans as innocent victims of the violent deaths of the Agents.[52] This outright violation of ahimsa stands at direct odds with the Buddhist ideal of compassion.

But why link knowledge so directly with violence? The filmmakers portray violence as redemptive,[53] and as absolutely essential to the success of the rebels. *The Matrix* steers sharply away at this point from the shared paradigms of Buddhism and Gnostic Christianity. The "reality" of the Matrix which requires that some humans must die as victims of salvific violence is not the ultimate reality to which Buddhism or Gnostic Christianity points. Neither the "stillness" of the *pleroma* nor the unchanging "nothingness" of *nirvana* are characterized by the dependence on technology and the use of force which so characterizes both of the worlds of the rebels in *The Matrix*.

The film's explicit association of knowledge with violence strongly implies that Neo and his comrades have not yet realized the ultimate reality. According to the worldviews of both Gnostic Christianity and Buddhism that the film evokes, the realization of ultimate reality involves a complete freedom from the material realm and offers peace of mind. The Wachowskis themselves acknowledge that it is "ironic that Morpheus and his crew are completely

dependent upon technology and computers, the very evils against which they are fighting."[54] Indeed, the film's very existence depends upon both technology's capabilities and Hollywood's hunger for violence. Negating itself, *The Matrix* teaches that *nirvana* is still beyond our reach.

IV. Concluding Remarks

Whether we view the film from a Gnostic Christian or Buddhist perspective, the overwhelming message seems to be, "Wake up!" The point is made explicit in the final song of the film, Wake Up!, by, appropriately, Rage Against the Machine. Gnosticism, Buddhism and the film all agree that ignorance enslaves us in an illusory material world and that liberation comes through enlightenment with the aid of a teacher or guide figure. However, when we ask the question, "To what do we awaken?", the film appears to diverge sharply from Gnosticism and Buddhism. Both of these traditions maintain that when humans awaken, they leave behind the material world. The Gnostic ascends at death to the pleroma, the divine plane of spiritual, non-material existence, and the enlightened one in Buddhism achieves *nirvana*, a state which cannot be described in language, but which is utterly non-material. By contrast, the "desert of the real," is a wholly material, technological world, in which robots grow humans for energy, Neo can learn martial arts in seconds through a socket inserted into the back of his brain, and technology battles technology (Nebuchadnezzar vs. AI, electromagnetic pulse vs. Sentinels). Moreover, the battle against the Matrix is itself made possible through technology - cell phones, computers, software training programs. "Waking up" in the film is leaving behind the Matrix and awakening to a dismal cyber-world, which is the real material world.

Or perhaps not. There are several cinematic clues in the scene of the construct loading program (represented by white space) that suggest that the "desert of the real" Morpheus shows Neo may not be the ultimate reality. After all, Morpheus, whose name is taken from the god of dreams, shows the "real" world to Neo, who never directly views the surface world himself. Rather, he sees it on a television bearing the logo "Deep Image." Throughout the film, reflections in mirrors and Morpheus's glasses, as well as images on television monitors point the viewer toward consideration of multiple levels of illusion.[55] As the camera zooms in to the picture on this particular television and the viewer "enters" the image, it "morphs" the way the surveillance screens do early in the film, indicating its unreality. In addition, the entire episode takes place while they stand in a construct loading program in which Neo is warned not to be tricked by appearances. Although sense perception is clearly not a reliable source for establishing reality, Morpheus himself admits that, "For a long time I wouldn't believe it, and then I saw the fields [of humans grown for energy] with my own eyes... And standing there, I came to realize the obviousness of the truth." We will have to await the sequels to find out whether "the desert of the real" is itself real.[56]

Even if the film series does not ultimately establish a complete rejection of the material realm, *The Matrix* as it stands still asserts the superiority of the human capacity for imagination and realization over the limited "intelligence" of technology. Whether stated in terms of matter/ spirit, body/ mind, hardware/ software or illusion/ truth, the ultimate message of *The Matrix* seems to be that there may be levels of metaphysical reality beyond what we can ordinarily perceive, and the film urges us to open ourselves to the possibility of awakening to them.

Frances Flannery-Dailey & Rachel Wagner

[Note: This essay originally appeared in *The Journal of Religion and FIlm*]

Endnotes

1. All unidentified quotes are from *The Matrix* (Warner Bros. release, 1999).

2. In an online chat with viewers of the DVD, the Wachowskis acknowledged that the Buddhist references in the film are purposeful. However, when asked "Have you ever been told that the Matrix has Gnostic overtones?", they gave a tantalizingly ambiguous reply: "Do you consider that to be a good thing?" From the Nov. 6, 1999 "Matrix Virtual Theatre," at "Wachowski chat"

3. Elaine Pagels notes that the similarities between Gnosticism and Buddhism have prompted some scholars to question their interdependence and to wonder whether "...if the names were changed, the 'living Buddha' appropriately could say what the *Gospel of Thomas* attributes to the living Jesus." Although intriguing, she rightly maintains that the evidence is inconclusive, since parallel traditions may emerge in different cultures without direct influence. Elaine Pagels, *The Gnostic Gospels*, (New York: Random House, 1979, repr. 1989), xx-xxi

4. James Ford recently explored other Buddhist elements in *The Matrix*, which he rightly calls a "modern myth," in his article "Buddhism, Christianity and *The Matrix*: The Dialectic of Myth-Making in Contemporary Cinema," for the *Journal of Religion and Film*, vol.4 no. 2. See also Conrad Ostwalt's focus on apocalyptic elements of the film in "*Armageddon* at the Millennial Dawn," JRF vol. 4, no. 1

5. A viewer asked the Wachowski brothers, "Your movie has many and varied connections to myths and philosophies, Judeo-Christian, Egyptian, Arthurian, and Platonic, just to name those I've noticed. How much of that was intentional?" They replied, "All of it" (Wachowski chat).

6. Feminists critics can rejoice when Trinity first reveals her name to Neo, as he pointedly responds, "The Trinity?... Jesus, I thought you were a man." Her quick reply: "Most men do."

7. The Wachowski brothers indicate that the names were "all chosen carefully, and all of them have multiple meanings," and also note this applies to the numbers as well (Wachowski chat).

8. In a recent interview in *Time*, the Wachowskis refer to Nebuchadnezzar in this Danielic context, (www.time.com/time/magazine/article/0,9171,22971,00.html , "Popular Metaphysics," by Richard Corliss, Time, April 19, 1999 Vol. 153, no. 15). Nebuchadnezzar is also the Babylonian king who destroyed the Jerusalem Temple in 586 B.C.E., and who exiled the elite of Judean society to Babylon. Did the Wachowski brothers also intend the reference to point to the crew's "exile" from Zion or from the surface world?

9. The film also suggests Zion is heaven, such as when Tank says, "If the war was over tomorrow, Zion is where the party would be," evoking the traditional Christian schema of an apocalypse followed by life in heaven or paradise. Ironically, the film locates Zion "underground, near the Earth's core, where it is still warm," which would seem to be a cinematic code for hell. Is this a clue that Zion is not the "heaven" we are led to believe it is?

10. Neo's apartment number is 101, symbolizing both computer code (written in 1s and 0s) and his role as "the One." Near the end of the film, 303 is the number of the apartment that he

enters and exits in his death / resurrection scene, evoking the Trinity. This in turn raises questions about the character of Trinity's relationship to Neo in terms of her cinematic construction as divinity.

11. The traitor Cypher, who represents Judas Iscariot, among other figures, ironically says to Neo, "Man, you scared the B'Jesus outta me."

12. We would like to thank Donna Bowman, with whom we initially explored the Gnostic elements of *The Matrix* during a public lecture on film at Hendrix College in 2000.

13. Gnosticism may have had its origins in Judaism, despite its denigration of the Israelite God, but the issue is complex and still debated within scholarly circles. It is clear, however, that Gnostic Christianity flourished from at least the 2nd - 5th c. C.E., with its own scriptures, and most likely also its own distinctive rituals, entrance requirements and a creation story. See Gershom Scholem, *Jewish Gnosticism, Merkabah Mysticism, and Talmudic Tradition* (New York: Jewish Theological Seminary of America, 1960), Elaine Pagels, *The Gnostic Gospels* (New York: Vintage Books, 1979, repr. 1989), Bentley Layton, *The Gnostic Scriptures* (New York: Doubleday, 1995), Kurt Rudolph, *Gnosis: The Nature and History of Gnosticism* (San Francisco: HarperSanFrancisco, 1987).

14. This corpus lay dormant for nearly 2000 years until its discovery in 1945 in Nag Hammadi, Egypt. The complete collection of texts may be found in James M. Robinson, ed. *The Nag Hammadi Library*, revised edition, (New York: HarperCollins, 1990; reprint of original Brill edition, 1978). These documents are also available on-line at The Nag Hammadi Library Section of The Gnostic Society Library.

15. Gnostic texts are cryptic, and no single text clearly explains this myth from beginning to end. The literature presupposes familiarity with the myth, which must be reconstructed by modern readers. The version of the myth presented here relies on such texts as *Gospel of Truth, Apocryphon of John, On the Origin of the World* and *Gospel of Thomas*. See *The Nag Hammadi Library*, pp. 38-51, 104-123, 124-138, 170-189.

16. Since the divine beings are composed only of spiritual substances and not matter, there are no physical gender differences among the beings.

17. Depending on the text, a plethora of divine beings populate the pleroma, many with Jewish, Christian or philosophical names, e.g. the Spirit, forethought, thought, foreknowledge, indestructibility, truth, Christ, Autogenes, understanding, grace, perception, Pigera-Adamas (*Apocryphon of John*).

18. Humanity's characterization also resonates with the Tower of Babel story in Genesis 11:1-9; in both we admire the work of our own hands.

19. The bulk of the following excerpt from the Gnostic "Gospel of Truth" might just as well be taken from the scenes in *The Matrix* in which Morpheus explains the nature of reality to Neo: Thus they [humans] were *ignorant* of the Father, he being the one whom they did not *see*... there were many *illusions* at work... and (there were) empty fictions, *as if they were sunk in sleep and found themselves in disturbing dreams*. Either (there is) a place to which they are fleeing, or without strength they come (from) having chased after others, or they are involved in striking blows, or they are receiving blows themselves, or they have fallen from high places, *or they take off into the air though they do not even have wings*. Again, sometimes (it is as) if people were murdering them, though there is no one even pursuing them, or they themselves are killing their neighbors...(but) *When those who are going through all these things wake up,* they see nothing, they who were in the midst of all these disturbances, for they are nothing. Such is the way of those who have cast ignorance aside from them like sleep, not esteeming it

as anything, nor do they esteem its works as solid things either, but *they leave them behind like a dream in the night*... This is the way each one has acted, as though asleep at the time when he was ignorant. And this is the way he has [come to knowledge], as if he had awakened.
(*Gospel of Truth*, 29-30)

20. This is perhaps most evident in the subway fight between Neo and Agent Smith. At a point in the film when Morpheus says of Neo, "He is just beginning to believe," Agent Smith calls him "Mr. Anderson," and while fighting he replies, "My name is Neo." The Wachowskis confirm this interpretation when they state "Neo is Thomas Anderson's potential self" (Wachowski chat).

21. This twin tradition was especially popular in Syrian Christianity. See also Pagels, p. xxi, where she wonders if the tradition that Thomas, Jesus' twin, went to India points to any historical connection between Buddhism and Hinduism on the one hand and with Gnosticism on the other.

22. See the online chat with the special effects creators in the "Matrix Virtual Theater" from March 23, 2000.

23. *Nag Hammadi Library*, pp. 490-500. Compare the Gnostic idea of stillness with these Buddhist sayings from the *Dhammapada*: "The bhikku [monk], who abides in loving-kindness, who is delighted in the Teaching of the Buddha, attains the State of Calm, the happiness of stilling the conditioned things" and "Calm is the thought, calm the word and deed of him who, rightly knowing, is wholly freed, perfectly peaceful and equipoised. " Quoted in Walpola Sri Rahula, *What the Buddha Taught* (New York: Grove Weidenfeld, 1974) p.128, 136.

24. See *Nag Hammadi Library*, pp. 256-59. We are grateful to Brock Bakke for the initial equation of agents with archons.

25. In Gnosticism "Mind" or the Greek "nous" is a deity, such as in the text "Thunder, Perfect Mind," *Nag Hammadi Library*, 295-303.

26. Note that as Morpheus and Neo enter the elevator of the apartment building of the Oracle, images of "seeing" symbolize prophecy and knowledge: a blind man (evoking blind prophets such as Tiresias) sits in the lobby beneath some graffiti depicting a pair of eyes. Interestingly, the Oracle - a sibyl / seer - wears glasses to look at Neo's palm.

27. Note too the metonymic use of color to convey this dualism: black and white clothing, floors, furniture, etc.

28. Ostwalt, "Armageddon" in JRF Vol. 4, no. 1. The parallel with apocalypticism does not work quite as well as one with Gnosticism because like Gnosticism, the film understands salvation to be individual (rather than collective and occurring all at once), to be attained through knowledge, and most importantly to entail leaving behind the material Earth (that is, not resulting in a kingdom of God made manifest on the Earth).

29. In its description in the original screenplay, the Temple of Zion evokes both the Oracle of Delphi (three legged stool, priestesses) and the Jerusalem Temple (polished marble, empty throne which is the mercy seat or throne of the invisible God).

30. A viewer asked the Wachowski brothers, "What is the role or {sic} faith in the movie? Faith in oneself first and foremost – or in something else?" They answered, "Hmmmm...that is a tough question! Faith in one's self, how's that for an answer?" This reply hardly settles the issue (Wachowski chat).

31. Specifically, these humans are Neo (the Gnostic Redeemer / Messiah) and Morpheus and Trinity, both of whom are named for gods. As a godhead, this trio does not quite make sense in terms of traditional Christianity. However, the trio is quite interesting in the context of Gnosticism, which portrays God as Father, Mother and Son, a trinity in which the Holy Spirit is identified as female, e.g. *Apocryphon of John* 2:9-14. For further reading on female divinities in Gnosticism, see Pagels, pp. 48-69.

32. The brothers explain, "There's something uniquely interesting about Buddhism and mathematics, particularly about quantum physics, and where they meet. That has fascinated us for a long time" (Wachowski chat). In the Time interview with Richard Corliss (see note 8), Larry Wachowski adds that they became fascinated "by the idea that math and theology are almost the same. They begin with a supposition you can derive a whole host of laws or rules from. And when you take all of them to the infinity point, you wind up at the same place: these unanswerable mysteries really become about personal perception. Neo's journey is affected by all these rules, all these people trying to tell him what the truth is. He doesn't accept anything until he gets to his own end point, his own rebirth." The film's presentation of the Matrix as a corporate network of human conceptions (or samsara) which are translated into software codes that reinforce one another illustrates this close relationship.

33. Stupa: a hemispherical or cylindrical mound or tower serving as a Buddhist shrine.

34. Of course, the most transparent reference to Buddhist ideas occurs in the waiting room at the Oracle's apartment, where Neo is introduced to the "Potentials." The screenplay describes the waiting room as "at once like a Buddhist temple and a kindergarten class." One of the children, clad in the garb of a Buddhist monk, explains to Neo the nature of ultimate reality: "There is no spoon." One cannot help wondering if this dictum only holds within the Matrix or if there is in fact "no spoon" even in the real world beyond it.

35. *Samyutta-nikaya* IV, 54. In Edward Conze, ed. *Buddhist Texts Through the Ages* (New York: Philosophical Library, 1954), p. 91.

36. *Samyutta-nikaya* II, 64-65. Ibid.

37. The entire process depends upon human ignorance, so that almost all who are born into the Matrix are doomed to be born, to die, and to re-enter the cycle again. When asked about the film's depiction of the liquefaction of humans, the Wachowskis reply that this black ooze is "what they feed the people in the pods, the dead people are liquefied and fed to the living people in the pods." Tongue in Buddhist cheek, the brothers explain this re-embodiment: "Always recycle! It's a statement on recycling."(Wachowski Chat) Even in the "real world" beyond the Matrix, the human plight is depicted as a relative and inter-dependent cycle of birth, death, and "recycling."

38. (Ed. Note: This clip can be viewed here. (Hit your back button to return to this essay.))

39. This dialogue also points to the "reality" (or the "Matrix") which we ourselves inhabit. In our world, and in the world of Joe Pantoliano, he is an actor. Therefore, the world of which both the actor Joe Pantoliano and we are now a part may be seen as the "Matrix" into which he has been successfully re-inserted, and thus the film itself may be seen as a part of the software program of our own "Matrix." The argument, of course, is seductively circular.

40. Take, for example, this quote from the *Sabbasava-sutta*: "A bhikku [monk], considering wisely, lives with his eyes restrained . . . Considering wisely, he lives with his ears restrained . . . with his nose restrained . . . with his tongue . . . with his body . . . with his mind restrained . . . a bhikku, considering wisely, makes use of his robes -- only to keep off cold, to keep off heat . . and to cover himself decently. Considering wisely, he makes use of food – neither for

pleasure nor for excess . . . but only to support and sustain this body . . ." (Quoted in Rahula 103).

41. James Ford has argued that the film embodies in particular the Yogacara school of Buddhism. Instead of pointing to that which is absolutely different than the world as *nirvana*, Yogacarins point to the world itself, and through the processes enacted in meditation, come to the realization that "all things and thought are but Mind-only. The basis of all our illusions consists in that we regard the objectifications of our own mind as a world independent of that mind, which is really its source and substance" (Edward Conze, *Buddhism*. New York: Philosophical Library, 1959), p. 167. The Matrix exists only in the minds of the human beings which inhabit it, so that in *The Matrix*, as in Yogacara, "The external world is really Mind itself" (p. 168). Yet a problem arises when one realizes that for the Yogacara school, the Mind is the ultimate reality, and therefore *samsara* and *nirvana* become identified. By contrast, the film insists on a *distinction* between *samsara* (the Matrix) and *nirvana* (that which lies beyond it). Because *The Matrix* maintains a duality between the Matrix and the realm beyond it, Yogacara is of limited help in making sense of the Buddhist elements in the film, nor is it helpful in supporting the idea that beyond the Matrix and beyond the Nebuchadnezzar there is an ultimate reality not yet realized by humans (see note 4).

42. According to Theravada teachings, *arhat* ("Worthy One") is a title applied to those who achieve enlightenment. Because, according to Theravada beliefs, enlightenment can only be achieved through individual effort, an *arhat* is of limited aid in helping those not yet enlightened and so would not necessarily choose to re-enter *samsara* to aid others still enmeshed within it.

43. Rahula, p. 2.

44. Quoted in Rahula, 135.

45. Quoted in Rahula, 133.

46. A bodhisattva is one who postpones final entry into *nirvana* and willingly re-enters or remains in *samsara* in order to guide others along the path to enlightenment. The Buddha's compassion serves as their primary model for Mahayana Buddhists, since they point out that he too remained in *samsara* in order to help others achieve enlightenment through his teachings and example.

47. The screenplay describes Neo as "floating in a womb-red amnion" in the power plant.

48. In the screenplay, Trinity does not kiss him but instead "pounds on his chest," precipitating his resuscitation. The screenplay states directly: "It is a miracle." This fourth "life" can be viewed as the one to which the Oracle refers in her predictions that Neo was "waiting for something" and that he might be ready in his "next life, maybe." This certainly appears to be the case, since Neo rises from the dead and defeats the Agents.

49. These four "lives" suggest that Neo is nothing other than "the One" foretold by the oracle, the reincarnation of the first "enlightened one," or Buddha, who "had the ability to change whatever he wanted, to remake the Matrix as he saw fit." Buddhist teaching allows that those who have been enlightened are endowed with magical powers, since they recognize the world as illusory and so can manipulate it at will. Yet supernatural powers are incidental to the primary goal, which is explained in the very first sermon spoken by the Buddha: "The Noble Truth of the cessation of suffering is this: It is the complete cessation of that very thirst, giving it up, renouncing it, emancipating oneself from it, detaching oneself from it"(*Dhammacakkappavattana-sutta*. Quoted in Rahula, 93.)

50. *Buddhacarita* 1:65. E. B. Cowell, trans., *Buddhist Mahayana Texts, Sacred Books of the East*, vol. 49 (Oxford: Oxford University Press, 1894).

51. See, for example, in the *Dhammapada*: "Of death are all afraid. Having made oneself the example, one should neither slay nor cause to slay" (Verse 129) (*Dhammapada*, trans. John Ross Carter and Mahinda Palihawadana. New York: Oxford University Press, 1987), p. 35.

52. The idea that violence as salvific is made explicit by the writers. Whereas they *could* have chosen to present the "deaths" of the Agents as of the same illusory quality as other elements within the software program, instead, they choose to depict *actual* humans *really* dying through the inhabitation of their "bodies" by the Agents. This addition is completely unnecessary to the overall plot line; indeed, the "violence" which takes place in the Hotel could still be portrayed, with the reassuring belief that any "deaths" which occur there are simply computer blips. The fact that the writers so purposefully insist that actual human beings die (i.e. die also within the power plant) while serving as involuntary "vessels" for the Agents strongly argues for *The Matrix*'s direct association of violence with the knowledge required for salvation.

53. See the article by Bryan P. Stone, "Religion and Violence in Popular Film," JRF Vol. 3, no. 1.

54. When asked whether this irony was intentional, the Wachowskis reply abruptly but enthusiastically "Yes!" (Wachowski chat).

55. This is especially true in the "red pill / blue pill" scene where Neo first meets Morpheus, and Neo is reflected differently in each lens of Morpheus's glasses. The Wachowskis note that one reflection represents Thomas Anderson, and one represents Neo (Wachowski chat).

56. A viewer asked the pertinent question of the Wachowskis: "Do you believe that our world is in some way similar to *The Matrix*, that there is a larger world outside of this existence?" They replied: "That is a larger question than you actually might think. We think the most important sort of fiction attempts to answer some of the big questions. One of the things that we had talked about when we first had the idea of *The Matrix* was an idea that I believe philosophy and religion and mathematics all try to answer. Which is, a reconciling between a natural world and another world that is perceived by our intellect" (Wachowski chat).

What's So Bad about Living in The Matrix?

James Pryor

There's a natural, simple thought that the movie *The Matrix* encourages. This is that there's something bad about being inside the Matrix. That is, there's an important respect in which people inside the Matrix are worse off than people outside it. Of course, most people inside the Matrix are ignorant of the fact that they're in this bad situation. They falsely believe they're in the good situation. Despite that, they are still worse off than people who *really are* in the good situation.

I said this is a natural, simple thought. When we look more closely, though, this natural, simple thought starts to get very complicated and unclear. Many questions arise.

First question: *Who* is the Matrix supposed to be bad for? Is life inside the Matrix only bad for people like Trinity and Neo who have experienced life outside? Or is it also bad for all the ordinary Joes who've never been outside, and have no clue that their present lives are rife with illusion? The movie does seem to suggest that there's something bad about life in the Matrix even for these ordinary Joes. It may be *difficult* to face up to the grim realities outside the Matrix, but the movie does present this as a choice worth making. It encourages the viewer to sympathize with Neo's choice to take the red pill. The character Cypher who chooses to reinsert himself into the Matrix is not portrayed very sympathetically. And at the end of the movie, Neo seems to be

embarking on a crusade to free more people from the Matrix.

What do you think? If *you* had the power to free people from the Matrix, would you use that power? We can assume that these people's minds are "ready," that is, they can survive being extracted from the Matrix without going insane. But let's suppose that once you freed them, they did not have the option of going back. Do you think they'd be better off outside? Would you free them? Do you think they'd thank you?

Or do you side with Cypher? Do you think that life inside the Matrix isn't all that bad—especially if your enjoyment of it isn't spoiled by the knowledge that it's all a machine-managed construct?

Second question: Does it matter who's running the Matrix, and why? In the movie, the machines are using the Matrix to keep us docile so that they can use us as a source of energy. In effect, we're their cattle. But what if we weren't at war with the machines? What if the machines' purposes were purely benevolent and philanthropic? What if they created the Matrix because they thought that our lives would be more pleasant in that virtual world than in the harsher real world? (Iakovos Vasiliou discusses a scenario like this in his essay.) Or what if we *defeated* the machines, took over the Matrix machinery ourselves, and then chose to plug ourselves back in because life inside was more fun? Would these differences make a difference to whether you regard life inside the Matrix as bad? Or to how bad you regard it?

In his third essay, Christopher Grau discusses Robert Nozick's "experience machine." Nozick thinks that there are things we value in life that we'd be losing out on if we plugged into an experience machine. He thinks there are things we lose out on even if the operators' intentions are benevolent and we

plug in of our own free choice. Do you think that's right? Would you say the same thing about the Matrix?

Our answers to these questions will be useful guides as we try to determine what it is about *the movie's version* of the Matrix that makes us squeamish.

II

In order to figure out what's so bad about being in the Matrix, it will help to do some conceptual ground-clearing.

When they think about scenarios like the Matrix, some people have the thought:

> If in every respect it seems to you that you're in the good situation, doesn't that make it *true*—at least, true for you—that *you are* in the good situation?

This line of thought is never fully endorsed in the movie, but the characters do sometimes flirt with it. Consider the conversation Neo and Morpheus have in the Construct:

> Neo: This isn't real...
>
> 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...

Consider Cypher's final conversation with Trinity:

> Cypher: ...If I had to choose between that and the Matrix...I choose the Matrix.
>
> Trinity: The Matrix isn't real.

> Cypher: I disagree, Trinity. I think the Matrix can be more real than this world...

Are the claims that Morpheus and Cypher are making here right? Is the world that Trinity and Cypher experience and seem to interact with when they're inside the Matrix just as real (or more real?) than the world outside?

The standard view is "no," the Matrix world is in some important sense less real. As Morpheus goes on to say, the Matrix is "a dream world." The characters are just experiencing a "neural interactive simulation" of eating steak, jumping between buildings, dodging bullets, and so on. As Neo says when he's on the way to visit the Oracle, "I have these memories from my life. None of them happened." In fact, he never has eaten steak, and never will. It just seems to him that he has.

And presumably that's how things would be even if no one ever *discovered* that it was so; even if no one ever *figured out* that the Matrix was just a "dream world."

Philosophers would express this standard view by saying that facts like:

- whether you've ever eaten steak
- whether you've ever jumped between buildings
- whether your eyes have ever been open

and so on are all *objective facts*, facts that are true (or false) independently of what anybody believes or knows about them, or has evidence for believing. The mere fact that *it seems to you* that you're jumping between buildings doesn't *make it true* that there really are any buildings there.

Some people get uneasy with this talk about "objective facts." They say:

> Well, what's true for me might be different than what's true for you. When *I'm* in the Matrix it really is *true for me* that I'm eating steak and so on. That might not be true for you, but it is true for me.

Let's try to figure out what this means.

Some of the time, people use expressions like "true for me" in a way that doesn't conflict with the view that the facts in question are objective.

For instance, all that some people mean by saying that something is "true for them" is that *they believe it to be true*. When you're in the Matrix you do believe that you're eating steak; so in this sense it will be "true for you" that you're eating steak. And what *you* believe to be true will often be different from what *I* believe to be true; so in this sense something could be "true for you" but "false for me." When a philosopher says that it's an objective fact whether or not you've ever eaten steak, she's not disputing any of this. She accepts that you and I *may disagree* about whether you've ever eaten steak. She's not even claiming to know who's right. She may be ignorant or mistaken about your past dietary habits, and she knows this. You may have better evidence than she, and she knows this too. All she's claiming is that *there is* a fact of the matter about whether you've eaten steak—regardless of whether you or she or anybody else knows what that fact is, or has any beliefs about it. And this fact is an objective one. If it happens to be true that you've eaten steak, then it's true, period. It's not "true for you" but "false for me." What you and I *believe*, and *who's got better evidence* for their belief, are further separate questions.

Usually when two people disagree about some matter, they agree that the fact

they're disputing is an objective one. They agree that one of them is right and the other wrong. They just disagree about who. For some matters, like ethical and artistic matters, this is less clear. It is philosophically controversial whether ethical and artistic truths are objective, and whether the same truths hold for everyone. But for our present discussion, we can set those controversies aside, and just concentrate on more prosaic and mundane matters, like whether you've ever eaten steak, whether your eyes have ever been open, and so on. For matters of this sort, we'd expect there to be only one single common truth, not one truth for you and a different truth for me.

Now, sometimes we speak incompletely. For example, we'll say that a kitchen gadget is useful, when we really mean that it's useful *for certain purposes*. It may be useful for cutting hard-boiled eggs but useless for cutting tomatoes or cheese. We'll say that the cut of certain suits makes them fit better, when we really mean that it makes them fit *certain people* better. It doesn't make them fit people with unusual body shapes better. And so on. In cases like this, if one way of completing the claim is natural when we're talking about you, and another way when we're talking about me, then we might be tempted to talk of the claim's being "true for you" but "false for me." For instance, suppose you're cutting eggs for a salad and I'm cutting the tomatoes. We're each using the same kitchen gadget, you with good results and me with frustrating results. If you say "This kitchen gadget is useful," I might respond "That may be true for you, but it's not true for me." There's no conflict here with the view that facts about usefulness are objective. Really there are *several* facts here:

•The gadget is useful for cutting eggs.
•The gadget is not very useful for cutting tomatoes.

•The gadget is more useful for you than it is for me (because you're cutting eggs and I'm cutting tomatoes.)

And so on. It's perfectly possible to regard all these facts as objective. That is, if any of them are true, then they're true, period. It won't be "true for you" that the gadget is more useful for you than it is for me, but "false for me." And neither will my *thinking that* the gadget is useless for cutting tomatoes make it so. I can be mistaken about how useful the gadget is. (Perhaps I'm not using it properly.) Similarly, if your new Armani suit doesn't fit you very well, then it doesn't fit you, even if we both somehow convince ourselves that it does fit.

So the ways of talking about things being "true for me" etc. that we've considered so far don't conflict with the view that the facts we're dealing with are objective.

People who dislike objective facts want to say something stronger. They want to say *it really is true* for the characters inside the Matrix that they've eaten steak. They're not just making a claim about what those characters *think* is true. When those characters think to themselves, "I've eaten steak hundreds of times, and so has my friend Neo," *what they're thinking really is supposed to be true*. At least for them. For Neo and Trinity and others it may not be true.

One way to flesh this idea out is with a philosophical theory called *verificationism*. (Sometimes this theory is called *anti-realism*.) If you're a verificationist about certain kinds of fact, then you reject the idea that those facts are objective. For example, a verificationist about height would say that *how tall you are* depends on *what evidence there is* about how tall you are. It's impossible for all the evidence to point one way, but the facts about your

height to be otherwise. The facts have to be *constrained by* the evidence. Sure, the verificationist will say, people sometimes make mistakes about their height. They sometimes have false beliefs. But those mistakes have to be in principle *discoverable* and *correctable*. It doesn't make sense to talk about a situation where everybody is permanently and irremediably mistaken about your height, where the "real facts" are so well-concealed that no one will be able to ferret them out. If the "real facts" are so well-concealed, says the verificationist, then they cease being facts at all. The only height you can have is a height that it's in principle discoverable or *verifiable* that you have. (Hence the name "verificationism.")

When we're discussing the Matrix and examples like it in my undergraduate classes, and students start talking about things being "true for" them, but "false for" other people, they're usually trying to sign onto some kind of verificationism. They'll say things like this:

> If all my evidence says that there is a tall mountain there, then in my personal picture of the world *there is* a tall mountain there. That's all it can *mean, for me*, to say that there's a tall mountain there. The mountain really is there, for me, so long as it appears real, and fits my conception of a tall mountain.

I'm always surprised to hear students voicing approval for this view. It's a pretty strange conception of reality. Some philosophers do defend the view. But I'd be really surprised if 30% of my university students really did think this is the way the world is. As a group, they don't usually tend to hold strange conceptions of reality; I don't find 30% of them believing in astrology or body-snatching aliens, for instance.

Mount Everest is 8,850 meters tall. Most of us think that Mt. Everest had this

height well before there were any human beings, and that it would still have this height even if no human beings or other thinking subjects had ever existed. But it's not clear that a verificationist is entitled to say things like that. If there had never been any thinking subjects, then there wouldn't have been anybody who could have *had evidence that* Mt. Everest existed. So according to the verificationist, then, there wouldn't have been anybody *for* whom it was true that Mt. Everest is 8,850 m tall. It looks like the verificationist has to deny that Mt. Everest would still have been 8,850 m tall, even in situations where no thinking subjects had ever existed. This is what makes verificationism such a strange view.

Perhaps the verificationist will respond: Granted, in the situation we're envisaging, *nobody actually has* evidence that Mt. Everest is 8850 m tall. But the evidence *is still available*. (Mt. Everest will cast shadows of certain lengths at certain times of the day, and so on.) And if people had existed, they could have gathered and used that evidence. Maybe that's enough to make it true that Mt. Everest is still 8,850 m tall in the situation we're envisaging.

Things get tricky here. For instance, it's not clear that the verificationist is entitled to say that Mt. Everest *would still cast those shadows*, even if no observers had existed. But rather than pursuing these tricky details, let's instead think about examples where the relevant evidence isn't even *available*.

The usual varieties of verificationism say that for there to be a 8,850 m tall mountain, it has to be *publicly verifiable* that the mountain exists and is 8,850 m tall. That is, there has to be evidence that *somebody somewhere* could acquire that demonstrates that it is 8,850 m tall. A different version of the view would focus instead on what *I myself* am able to verify. This view might say

that it's "true for me" that the mountain is 8,850 m tall only if *I* could verify that it's 8,850 m tall. It'd be "true *for you*" that it's 8,850 m tall only if *you* could verify that it's 8,850 m tall. And so on. We can call this second version of the view "personal verificationism," since it says that what's true—well, true for me—always depends on what I myself would be able to verify. If there's some fact that will forever be concealed from me, then it's not really a fact; at least, not a fact "for me." It may be a fact for other people, but that's a separate issue.

When professional philosophers discuss verificationism, they usually have the public version in mind. And the two versions do share many of the same features—and problems. However, I'm just going to talk about the personal version of the view. I think that people who aren't professional philosophers, like the students in my undergraduate classes, usually find the personal version more natural and attractive.

What does it mean to say that certain evidence is "available" or "unavailable"? One way of drawing this line would make it turn on whether you can obtain the evidence through your own active efforts: e.g., are there tests you can run that would give you the evidence you need? Or you might have a more liberal conception of what it is for evidence to be "available." On this more liberal conception, evidence will count as "available" even if it could just happen to fall into your lap, by chance. It doesn't have to be in your power to make the evidence appear.

Let's think about someone for whom evidence is unavailable even on this more liberal conception of "available." Suppose there's a character in *The Matrix* that it's impossible for Morpheus to "waken." Maybe this character believes in the

"dream world" too strongly, and would just go insane and die if the "dream" ever started to unravel. Let's call this character Jeremy. According to the standard view, Jeremy has many false beliefs about his surroundings. He believes that he goes to work everyday on the 40th floor of an office building, that the sun streams into his office most mornings, that he often eats steak for dinner, and so on. All of these beliefs are false. In fact, there are no office buildings anymore; Jeremy has never seen the sun; he's never eaten steak; and he's spent his entire life in a small pod. But these are facts that Jeremy will never know. What's more, he's incapable of knowing them. If Morpheus told Jeremy the truth, Jeremy wouldn't believe him; and if Morpheus tried to *show* Jeremy the truth, Jeremy would go insane and die. So there are many truths about Jeremy's life that Jeremy will never be able to know.

That's what the standard view says. According to the verificationist, though, if it's impossible for Jeremy to know something, then that thing can't really be a "truth" about Jeremy's life. At least, it won't be a truth *for Jeremy*. What's true *for Jeremy* is that he really does work on the 40th floor of an office building, and so on. And this doesn't just mean that *Jeremy thinks* he works on the 40th floor etc. It means *it really is a fact*—a fact for Jeremy—that he works on the 40th floor of an office building. It may not be true for Morpheus that Jeremy works on the 40th floor of an office building, but it is true for Jeremy.

What do you think? Does that sound plausible to you?

Let's think about the comings and goings of people in the past. According to the standard view, on a given evening in the past, these people will either have been at a party in New York, or they won't have been there. Suppose they were there. But today only a little bit of evidence remains that they were there.

Suppose you have it in your power to destroy that evidence, and manufacture evidence that they were elsewhere. Would you then have it in your power to change the past? That is what the character O'Brien in George Orwell's novel *1984* thinks:

> An oblong slip of newspaper had appeared between O'Brien's fingers. For perhaps five seconds it was within the angle of Winston's vision... It was another copy of the photograph of Jones, Aaronson, and Rutherford at the party function in New York, which he had chanced upon eleven years ago and promptly destroyed. For only an instant it was before his eyes, then it was out of sight again...
>
> "It exists!" he cried.
>
> "No," said O'Brien.
>
> He stepped across the room. There was a memory hole in the opposite wall. O'Brien lifted the grating. Unseen, the frail slip of paper was whirling away on the current of warm air; it was vanishing in a flash of flame. O'Brien turned away from the wall.
>
> "Ashes," he said. "Not even identifiable ashes. Dust. It does not exist. It never existed."
>
> "But it did exist! It does exist! It exists in memory. I remember it. You remember it."...
>
> O'Brien was looking down at him speculatively. More than ever he had the air of a teacher taking pains with a wayward but promising child.
>
> "There is a Party slogan dealing with the control of the past," he said. "Repeat it, if you please."
>
> "'Who controls the past controls the future: who controls the present controls the past,'" repeated Winston obediently.
>
> "'Who controls the present controls the past,'" said O'Brien, nodding his head with slow approval. "Is it your opinion, Winston, that the past has real existence?... Is there somewhere or other a place, a world of solid objects, where the past is still happening?"
>
> "No."
>
> "Then where does the past exist, if at all?"
>
> "In records. It is written down."
>
> "In records. And—?"

> "In the mind. In human memories."
>
> "In memory. Very well, then. We, the Party, control all records, and we control all memories. Then we control the past, do we not?"

Now, presumably O'Brien knows he's tampered with the evidence. So perhaps he can't change what's true *for him* about the past. But on the verificationist view, it does seem like he'd be able to change the past for other people.

What do you think? Does that sound plausible? Winston eventually comes to accept this view of reality. But to the reader it's supposed to sound like a lie.

What if the machines in *The Matrix* said to Neo and Morpheus, "Hey, why do you keep harping about this war between humans and machines? It never happened. At least, for all these people in their pods we're making it true that it never happened. Once we've removed every shred of evidence, and made it impossible for them to verify that there was a war between humans and machines, then *we really will have* changed the past for those people. They won't be *deceived*. Their past *really will* have happened the way it seems to them." Does that sound convincing? Or does it too sound like a lie?

What about facts for which there's simply no evidence either way? Morpheus says they don't know who struck first in the war between humans and machines. Maybe it's not important. And maybe the machines don't know either. Maybe all the evidence is lost. But presumably one of us *did* strike first. Presumably *there is* a fact about this, even if there's no evidence remaining. The verificationist has to deny this.

I hope all of this will make verificationism sound somewhat implausible to you. They aren't meant to be conclusive considerations. Philosophical discussions of

verificationism get very complicated. The verificationist has to overcome many technical difficulties: e.g., how to draw the line between evidence that's available and evidence that's not. How to explain when evidence enables us to verify a hypothesis and when it doesn't. Whether verificationism itself is something we can verify. We can't go into these issues. If you're still inclined towards verificationism, I hope you'll at least grant that the view does go against our common-sense conception of reality, and that as a result it requires careful supporting argument. If you're going to hold the view in good intellectual conscience, there are a lot of difficulties and objections that need to be overcome.

III

I propose we set verificationism aside at this point; and see whether doing so helps us get any closer to determining what it is about the Matrix that makes it seem bad.

So now we'll say *it is* an objective fact whether you work on the 40th floor of an office building. We'll grant that *it can seem to you* in every respect that you're in "the good" situation (outside the Matrix), without it's thereby *being true* that you're in that situation.

OK. But this doesn't yet tell us why being inside the Matrix should be *bad*. Why is it important to *really be* in the situation we're calling "good"? Why isn't it good enough for us that we *seem to be* in the "good" situation? Isn't *the experience or illusion* of being in the good situation already pretty good? Why should it make our lives any better to *really* be there? (Especially if, as in the movie, the way the *real* "good" situation is is much less pleasant than the way

things *seem* to be in the so-called "bad" situation.)

As Cypher says:

> You know, I know that this steak doesn't exist. I know that when I put it in my mouth, the Matrix is telling my brain that it is juicy and delicious. After nine years, you know what I realize?... Ignorance is bliss.

Would it really make Cypher's life any better if he were *really* eating steak? Is it *really eating* steak that we value, or just *the experience* of eating steak? Wouldn't most people be satisfied with the experience—especially if it's indistinguishable from the real activity? Recall our friend Jeremy who spends his whole life inside the Matrix. How much is he missing out on, just because he never *really* gets to eat a steak? We're granting that there are truths about Jeremy's life that he'll never be able to know. But it's not obvious yet that any of them are *truths he cares about*. Perhaps the only things that Jeremy, and most of us, really care about are what kinds of experiences we're going to have, now and in the future. As Cypher recognizes, people who are stuck in the Matrix can still do pretty well by that score.

As we saw before, Nozick thinks that most of us *wouldn't* choose to spend the rest of our lives plugged into an "experience machine." He thinks there are things we value in life over and above what experiences we have. For instance, we value *doing* certain things, and not merely having the illusion or experience of doing them.

I agree with Nozick. For *some* matters, I think we genuinely *do* care about more than just what experiences we end up having. It would be implausible to claim this is always so. With regard to eating steak, the experience probably *is* all that we really value. But I think we feel differently about other matters. I'm

going to try to persuade you that this is so, too.

Notice that what we're talking about here is the question: *What do* we actually value? Not the question: *What should* we value? Some readers may be willing to concede that we *should* care about more than our own experiences. (It's so selfish!) But it may appear that, as a matter of fact, our own experiences are all we really *do* care about—at least most of us. I'm going to argue that this isn't so. Most of us *do* in fact care about more than just what experiences we end up having.

There's a widely-held picture of human motivation that makes it difficult to see this. That picture goes like this. Ultimately, it says, everyone always acts for selfish motives. Whenever we do something on purpose, it's *our own* purpose that we're trying to achieve. We're always pursuing *our own* ends, and trying to satisfy *our own* desires. All that any of us are really after in life is getting more pleasant experiences for himself, and avoiding painful ones. Sometimes it may *seem* that we're doing things for other people's sake. For instance, we give money to charity, we buy presents for our children, we make sacrifices to please our spouses. But if you look closer, you'll see that even in cases like these, we're still always acting for selfish motives. We only do such things because it makes us feel good and noble to do them, and we like feeling noble. Or we do them because when people we care about are happy, that makes *us* happy too, and ultimately what we're after is that happiness for ourselves. Hence, since the only aim we have in life is just to have pleasant experiences, Nozick's experience machine gives us everything we want, and it would be foolish not to plug into it.

Now, I grant that *some* people may be as selfish as this picture says. But I

doubt that many people are. The picture rests on two confusions, and once we clear those confusions up, I think there's no longer reason to believe that the *only* thing that *any of us* ever aims for in life is to have pleasant experiences.

The first confusion is to equate "pursuing our own ends, and trying to satisfy our own desires" with "acting for a selfish motive." To call a motive or aim "selfish" isn't just to say that it's a motive or aim that I have. It says more than that. It says something about *the kind of motive it is*. If my motive is to make me better off, then my motive is a selfish one. If my motive is to make you better off, then my motive is not selfish. From the mere fact that I'm pursuing one of *my* motives, it doesn't follow that my motive is of the first sort, rather than the second.

Ah, you'll say, but if my aim is to make you better off, then when I achieve that aim, I'll feel good. And this good feeling is really what I'll have been trying to obtain all along.

This is the second confusion. It's true that often when we get what we want (though sadly not always), we feel good. It's easy to make the mistake of thinking that what we *really* wanted was that good feeling. But let's think about this a bit harder. *Why* should making someone else better off give me a good feeling? And how do I know that it will have that effect?

Consider two stories. In story A, you go to visit the Oracle, and in her waiting room you see a boy bending spoons and a girl levitating blocks. You feel this inexplicable and unpleasant itch. Someone suggests as a hypothesis that the itch would go away if you gave the girl a spoon too. So you do so, and your itch goes away.

In story B, you walk into the same room, and you don't like the fact that the girl has no spoon. You would like her to have a spoon too. So you take a spoon and give it to the girl, and you feel pleased with the result.

In story A, your aim was to make yourself feel better, and giving the spoon to the girl was just a means to that end. It took experience and guesswork to figure out what would make you feel better in that way. In story B, on the other hand, no guesswork or experience seemed to be necessary. Here you were in a position to straightforwardly predict what would bring you pleasure. You could predict that because you had an aim *other than* making yourself feel better, you knew what that aim was, and usually you feel pleased when you get what you want. Your aim was to give a spoon to the girl. Your feeling of pleasure was a *consequence* or *side-effect* of achieving that aim. The pleasure is not what you were primarily aiming at; rather, it came about *because you achieved* what you were primarily aiming at. Don't mistake *what you're aiming at* with *what happens as a result of your getting* what you're aiming at.

Most often, when we do things to make other people better off, we're in a situation like the one in story B. Our pleasure isn't some unexplained effect of our actions, and what we're primarily trying to achieve. Our pleasure comes about *because we got* what we were primarily trying to achieve; and this makes it understandable why it should come about when it does.

Once we're straight about this, I think there's no argument left that the only thing anyone ever aims for in life is to have pleasant experiences. Some people do aim for that, some of the time. But many cases of giving to charity, making sacrifices for one's spouse, and so on, are not done for the pleasure they bring

to oneself. There's something else that one is after, and pleasure is just a pleasant side-effect that sometimes comes along with getting the other things one is after.

Nozick said that most of us *do* value more than our own experiences, that there are things that we value that we'd miss out on if we plugged into the Matrix. I think Nozick is right. He's right about me, and he's probably right about you, too. We can easily find out. I've devised a little thought-experiment as a test.

Suppose I demonstrate to you that your friends and I are very good at keeping secrets. For instance, one day when Trinity isn't around, we all make lots of fun of her. We read her journal out loud and laugh really hard. We do ridiculous impersonations of her. And so on. It's hilarious. But of course we only do this behind Trinity's back. When she shows up, nobody giggles or snickers or anything like that. You're completely confident that we'll be able to keep our ridicule a secret from Trinity. She'll never know about it.

Suppose I also demonstrate to you that I am a powerful hypnotist. I can make people forget things, and once forgotten they never remember them. You're convinced that I have this power.

Now that you know all of that, I offer you a choice. Option 1 is I deposit $10 in your bank account, but then your friends and I will make fun of you behind your back, the way we made fun of Trinity. If you choose this option, then I will immediately use my hypnotic powers to make you forget about making the choice, being teased, and all that. From your point of view, it will seem that the bank made an error and now you have $10 more in your account than you

had before. So in terms of what experiences you will have, this option has no downside. You won't even have to suffer from *the expectation* of being secretly teased, because I'll make you forget the whole arrangement as soon as you make your choice.

Option 2 is we keep things as they are. I pay you nothing, and your friends are no more or less likely to make fun of you behind your back than they were before.

So which would you choose?

When I offer my students this choice, I find that at least 95% of them choose Option 2. They think that the teasing would be a bad thing, even though they'd never know it was going on.

If the teasing doesn't seem so bad, then change the example. Say that in Option 1, your lover is cheating on you, but you never know about it. Or say that we're torturing your mother, but you never know about it. In every version, your experiences are smooth and untroubled, plus you get a little extra money. Which option would you choose?

If you find Option 2 more attractive, then that's support for Nozick's claim. The experience machine wouldn't give you everything you value. Option 1 gives you no experiences of being teased. It gives you no evidence that your lover is cheating on you, or that your mother is being tortured. But you don't just want to *have experiences* of things going well for yourself and your mother. You value *really* not being teased, *really* having a faithful lover, and *really* having an untortured mother.

Now, we do have to *compare* what we'd get by plugging into the experience machine to what we'd get if we don't plug in. I've only been arguing that we'd miss out on *some* things we'd value if we plugged in. I haven't said that it would *never* be reasonable to plug in. In some cases, the good of being plugged in could outweigh the bad. If the real world is miserable and nasty enough, it may make sense to plug in. Perhaps for Cypher, the real world is too nasty. All I'm saying is that plugging in won't give us *everything* we want. Our experiences aren't *all* that we value. So *there is* some bad to plugging in. There may also be some good to plugging in. Dreams and immersive role-playing do give us *some* of the things we value in life. I'm just saying they don't give us everything. *Some* aspects of how the world *really is* are important to us.

I haven't been able to say yet *how* important, though. It's hard to know what the right balance point is. How bad does the real world have to be, before it makes sense to make Cypher's choice, and plug back into the blissful experience machine? This is a hard question. In part, it will depend on whether the Matrix or the experience machine involve any hidden costs. And this is something we haven't yet settled.

IV

Before we can determine what are the major costs of living inside the Matrix, we have to confront one last complication.

We said that for most people inside the Matrix, the experience of eating steak may be enough. We said they probably don't care about whether they've ever *really* eaten steak. Let's pause over this for a moment. What do these

characters *mean* by "eating steak"?

Suppose you grew up with a friend you called "Jiro." You didn't realize it, but that isn't really your friend's name, at least not the name his parents gave him. His name is really "Takeshi." "Jiro" is his uncle's name. But you got the names mixed up when you were little, and no one bothered to correct you. So all your life you've been saying "Jiro" to talk about Takeshi. Isn't it plausible then that in your mouth, "Jiro" now *means* Takeshi?

Similarly, Jeremy has grown up inside the Matrix program, and on various occasions he's interacted in certain ways with other parts of the Matrix program, ways he described as "eating steak." Now perhaps *all he means* by "eating steak" is just interacting in those certain way with the Matrix. *He's done that* many times. So perhaps he *really has* managed to eat steak on many occasions. At least, he's managed to do what *he* calls "eating steak." It's not clear that there's *anything more* that Jeremy *would like* to be doing, but isn't. Is there?

The philosophical issues here are fascinating, but they get complicated really fast. I myself think that for *some* of Jeremy's concepts, the story we just sketched may be right.

Interestingly, this doesn't seem to be the movie's own attitude. Recall what Cypher says:

> You know, I know that this steak doesn't exist.

And when Morpheus and Neo are fighting in the sparring program, Morpheus asks:

Do you think that's air you're breathing?

Cypher and Morpheus are both rejecting the view that the Matrix simulations *really provide* what they mean by "steak" and "air." That is, they're rejecting the view that *all they mean* by "steak" and "air" is interacting in certain ways with the Matrix program.

As I said, the philosophical issues here can get really complicated. One way to avoid these difficulties is to concentrate on what would be bad about living in the Matrix *for the first generation of Matrix inductees*: people who grew up outside the Matrix, and have just been freshly plugged in. Presumably what *they* mean by "eating steak" has to do with cow flesh, not with patterns in the Matrix simulation. Presumably what *they* mean by "air" is made up of nitrogen and oxygen, not 1s and 0s.

I want to try a different strategy. We can suppose we're talking about people who have spent all their lives so far inside the Matrix. I want to try to find something we value that goes beyond what experiences we're having, and where we can agree that the people inside the Matrix *really would value that same thing*. They wouldn't just value having some Matrix substitute. And yet this will be something that people inside the Matrix don't have. They only seem to have it.

If we can find something like that, then we'll have found something that really does deserve the name of "what's bad about living in the Matrix."

I can think of three possibilities.

The first has to do with certain kinds of scientific knowledge. I'd guess that physicists in the Matrix have some fundamentally false beliefs about the underlying make-up of their world, what the "laws of nature" are, and so on. For some people, figuring such matters out is important. They value learning the truth about those matters. But not everybody feels that way. For your average non-physicist, the possibility that we're mistaken about questions like these isn't going to provoke existential anxiety, or set them off on a crusade like the one Neo undertakes at the end of the movie.

The second candidate for being what's bad about living in the Matrix has to do with interpersonal relationships. One thing we place a lot of value on in life are our interactions with other people. Most of us want our friends' feelings to be genuine. For instance, it would be bad if the person who acts like your best friend really despises you. Even if you never found out about it. Most of us also want the important people in our lives to be *real*. We don't want them to be programming constructs, like Mouse's "Woman in Red." Perhaps for some people, programming constructs are enough. They may not care whether their friends and lovers really have an inner life of their own, and have their own thoughts and emotions, and genuine feelings towards them. It would be enough if their friends and lovers acted the their parts well. I think that for most of us, though, this would not be enough. Most of us really would like to have the real thing. It would suck if the children you devote so much love and attention to are really just parts of a computer program, and don't have any capacity to benefit from, or to appreciate, your efforts.

Here's another thought-experiment. Suppose that tomorrow we're going to wipe your memory clean and ship you off to a new colony. You'll be able to live a decent life there; you just won't have any memory of your past. Nor do you

get to take any of your money or personal belongings along with you. But today, before we wipe your memory clean, we allow you to spend the money you have left to arrange a nice life for yourself in the new colony. For instance, if you spend $1,000 we'll set it up so that the apartment you get there doesn't have cockroaches. And so on. How would you spend your money?

What if there were two options on the menu. If you choose Option 1, you'll get an extremely realistic set of friends and lovers in the new colony. You won't be able to distinguish them from the real thing. But really they'll just be empty shells animated by a (non-intelligent) computer program. They won't have any inner life of their own. (In the terminology of role-playing games, they'll be NPCs.) You know this now, but when you get to the colony you will have forgotten it. If you choose Option 2, you get friends and lovers who are real people.

Most people I know would choose Option 2, even if it were somewhat more expensive, and so kept them from buying other nice things for their new life. E.g., they'd choose Option 2 even if it meant they'd have to put up with more cockroaches.

So one thing that many of us value in life is that the other people we form emotional attachments to are real people, and that they care about us in the ways they seem to. In Nozick's experience machine, this seems to be lacking. His experience machine sounds like a one-person Matrix. You just get to enjoy your own experience script. You don't get to interact with other people. (See the discussion of "solitary Matrices" in Richard Hanley's essay.)

In the real Matrix, on the other hand, it seems like people *do* get to interact

with many other real human beings. So a lack of interpersonal relationships may be a bad thing about Nozick's machine, but it doesn't seem to be a bad thing about the Matrix we see in the movie.

I think our third candidate for what's bad about living in the Matrix is more apt. In the movie, humans in the Matrix are all slaves. They're not in charge of their own lives. They may be contented slaves, unaware of their chains, but they're slaves nonetheless. They have only a very limited ability to shape their own futures. As Morpheus puts it:

> What is the Matrix? Control. The Matrix is a computer-generated dream world, built to keep us under control...

Now—to me anyway—the most disturbing thing about this isn't that the machines are farming us for energy. We're not told enough about how the energy-farming works to make it seem very bad. Perhaps the machines are only taking energy we were making no use of, anyway. Perhaps the machines ensure that—except for the rare occasions when an Agent takes over your body and gets it killed—we live longer and healthier lives in the Matrix energy-farm than we would in the wild.

No, what seems awful about our enslavement in the Matrix is rather that *our enemies have so much control over what happens to us*. Suppose we discovered that a secret Nazi cabal were really running our government. Wouldn't that be awful? Suppose they're not actively causing any harm. Suppose that for the most part they'll keep the government running in ways we like. Many of us still wouldn't like it. We'd object to the mere fact of those old Nazis having so much power over us.

Similarly, the machines in the Matrix are our enemies. We've fought a brutal

war with them. Now they have immense power over us. As long as it suits their purposes, they'll manage our lives in ways we like. But many of us will still be disturbed by their having so much power over us. We want to be in control of our own futures.

According to Agent Smith, the Matrix was designed to simulate the end of the 20th century, because the machines have found that keeps their energy-farm running smoothly. Generations of us have now lived out their lives in the Matrix. So generations of us have all experienced life in this simulated end-of-the-20th-century. What happens when the simulation gets up to 2003? Do the machines erase our memories and reset everything back to 1980? The movie doesn't say. But presumably they do something like that. This means there are real limits to how much we can accomplish. If your ambitions in the Matrix are relatively small-scale, like opening a restaurant or becoming a famous actor, then you may very well be able to achieve them. But if your ambitions are larger—e.g., introducing some long-term social change—then whatever progress you make towards that goal will be wiped out when the simulation gets reset. Any long-term efforts of this sort would be an exercise an futility.

And what if our ambitions don't please the farmers? For instance, what if we are computer scientists working to create artificial intelligence? The machines would probably find it easiest to just keep sabotaging our attempts. After all, they wouldn't want us to re-enact the war between humans and machines, inside the Matrix. That would be bad for their crops. And they certainly wouldn't want us to create *benevolent* AIs, AIs who would figure out about the Matrix and fight on our side. So the machines will tinker with our history, and see to it that grand, noble ambitions of this sort never get realized.

Of course, they'll also see to it that none of our grander *baser* ambitions get

realized, either. They probably just disconnect or reprogram anyone who's hatching plans for mass genocide.

But if given the choice, I think most of us would like humans to be in charge of our own destiny. We don't want our long-term efforts to be futile. We don't want to be living out someone else's plan for our lives. Sure, there will always be *some* limits to what we can do. Very likely we'll never be able to vacation in the center of the sun. But we'd like to have as much control over our destiny as we can. We don't want other intelligent agents deciding such things for us. *Especially* when those agents' first priority is how well their energy-farms are doing; that may not correlate well with how well-off our lives and society are.

So it seems rotten if the machines control our fates and our civilization. One thing we place a lot of value on is being in charge of our own lives, not being someone else's slave or plaything. We want to be *politically free*.

And plausibly, what people mean by "political freedom" and "being in charge of our own lives" is the same inside the Matrix as outside it. We're not indifferent between the real thing and some Matrix simulation of it. We want to have *the real thing*. When we're inside the Matrix, we haven't got it. We just don't realize that we haven't got it.

So I think this is the best answer about what's so bad about living in the Matrix.

For me, at least, it's a surprising answer. The Matrix raises so many interesting metaphysical and epistemological issues. If you're of a philosophical bent, like me, then those issues will be intellectually compelling. But there's a difference between what we find intellectually compelling and what we place the most

value on in life. Intellectual matters will be only one value among many. For most of us, the worst thing about living in the Matrix would not be something metaphysical or epistemological. Rather, the worst thing would be something *political*. It would be the fact that *Life in the Matrix is a kind of Slavery*, of the sort of we've discussed.

I think *that* is what's really bad about living in the Matrix we see in the movie. *That* is what motivates Neo and Morpheus and Trinity to fight the machines, and try to free everyone they can.

If the Matrix *weren't* a kind of enslavement—and it still involved interacting with other real people—then maybe it wouldn't be so bad after all.

James Pryor

The Matrix as Metaphysics

David Chalmers

I. 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. And for now, I will ignore the complication 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.

II. 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 Irish 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 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.

III. 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 hypotheses, 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.

I will make the case by introducing each of the three parts of the Metaphysical Hypothesis separately. I will suggest that each of them is coherent, and cannot be

conclusively ruled out. And I will suggest that none of them is a skeptical hypothesis: even if they are true, most of our ordinary beliefs are still correct. The same goes for a combination of all three hypothesis. I will then argue that the Matrix Hypothesis hypothesis is equivalent to this combination.

(1) The Creation Hypothesis

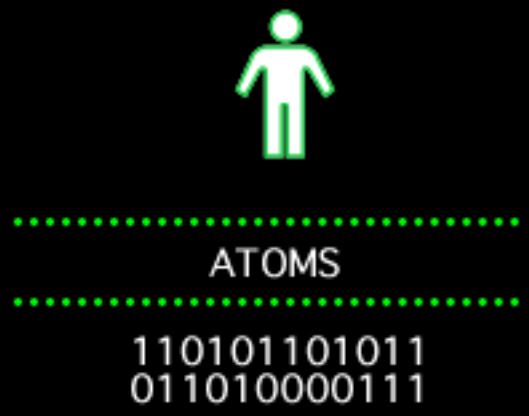
The Creation Hypothesis says: Physical space-time and its contents were created by beings outside physical space-time.

This is a familiar hypothesis. A version of it is believed by many people in our society, and perhaps by the majority of the people in the world. If one believes that God created the world, and if one believes that God is outside physical space-time, then one believes the Creation Hypothesis. One needn't believe in God to believe the Creation Hypothesis, though. Perhaps our world was created by a relatively ordinary being in the "next universe up", using the latest world-making technology in that universe. If so, the Creation Hypothesis is true.

I don't know whether the Creation Hypothesis is true. But I don't know for certain that it is false. The hypothesis is clearly coherent, and I cannot conclusively rule it out.

The Creation Hypothesis is not a skeptical hypothesis. Even if it is true, most of my ordinary beliefs are still true. I still have hands, I am still in Tucson, and so on. Perhaps a few of my beliefs will turn out false: if I am an atheist, for example, or if I believe all reality started with the Big Bang. But most of my everyday beliefs about the external world will remain intact.

(2) The Computational Hypothesis



The Computational Hypothesis says:
Microphysical processes throughout space-time are constituted by underlying computational processes.

The Computational Hypothesis says that physics as we know it not the fundamental level of reality. Just as chemical processes underlie biological processes, and microphysical processes underlie chemical processes, something underlies microphysical processes. Underneath the level of quarks and electrons and photons is a further level: the level of bits. These bits are governed by a computational algorithm, which at a higher-level produces the processes that we think of as fundamental particles, forces, and so on.

The Computational Hypothesis is not as widely believed as the Creation Hypothesis, but some people take it seriously. Most famously, Ed Fredkin has postulated that the universe is at bottom some sort of computer. More recently, Stephen Wolfram has taken up the idea in his book *A New Kind of Science*, suggesting that at the fundamental level, physical reality may be a sort of cellular automata, with interacting bits governed by simple rules. And some physicists have looked into the possibility that the laws of physics might be formulated

computationally, or could be seen as the consequence of certain computational principles.

One might worry that pure bits could not be the fundamental level of reality: a bit is just a 0 or a 1, and reality can't really be zeroes and ones. Or perhaps a bit is just a "pure difference" between two basic states, and there can't be a reality made up of pure differences. Rather, bits always have to be implemented by more basic states, such as voltages in a normal computer.

I don't know whether this objection is right. I don't think it's completely out of the question that there could be a universe of "pure bits". But this doesn't matter for present purposes. We can suppose that the computational level is itself constituted by an even more fundamental level, at which the computational processes are implemented. It doesn't matter for present purposes what that more fundamental level is. All that matters is that microphysical processes are constituted by computational processes, which are themselves constituted by more basic processes. From now on I will regard the Computational Hypothesis as saying this.

I don't know whether the Computational Hypothesis is correct. But again, I don't know that it is false. The hypothesis is coherent, if speculative, and I cannot conclusively rule it out.

The Computational Hypothesis is not a skeptical hypothesis. If it is true, there are still electrons and protons. On this picture, electrons and protons will be analogous to molecules: they are made up of something more basic, but they still exist. Similarly, if the Computational Hypothesis is true, there are still tables and chairs, and macroscopic reality still exists. It just turns out that their fundamental reality is a little different from what we thought.

The situation here is analogous to that with quantum mechanics or relativity. These may lead us to revise a few "metaphysical" beliefs about the external world: that the world is made of classical particles, or that there is absolute time. But most of our ordinary beliefs are left intact. Likewise, accepting the Computational Hypothesis may lead us to revise a few metaphysical beliefs: that electrons and protons are fundamental, for example. But most of our ordinary beliefs are unaffected.

(3) The Mind-Body Hypothesis

The Mind-Body Hypothesis says: My mind is (and has always been) constituted by processes outside physical space-time, and receives its perceptual inputs from and sends its outputs to processes in physical space-time.

The Mind-Body Hypothesis is also quite familiar, and quite widely believed. Descartes believed something like this: on his view, we have nonphysical minds that interact with our physical bodies. The hypothesis is less widely believed today than in Descartes' time, but there are still many people who accept the Mind-Body Hypothesis.

Whether or not the Mind-Body Hypothesis is true, it is certainly coherent. Even if contemporary science tends to suggest that the hypothesis is false, we cannot

rule it out conclusively.

The Mind-Body Hypothesis is not a skeptical hypothesis. Even if my mind is outside physical space-time, I still have a body, I am still in Tucson, and so on. At most, accepting this hypothesis would make us revise a few metaphysical beliefs about our minds. Our ordinary beliefs about external reality will remain largely intact.

(4) The Metaphysical Hypothesis

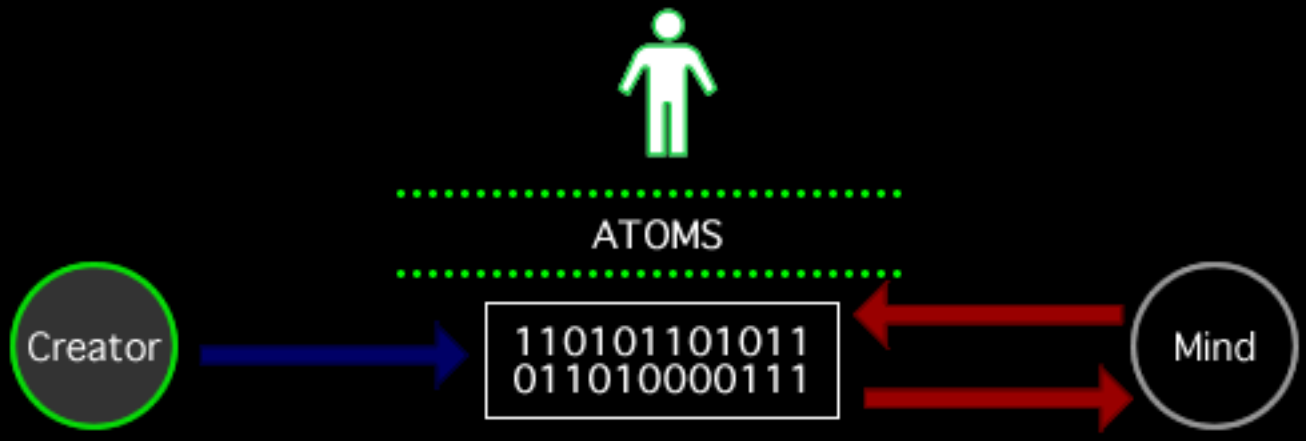
We can now put these hypotheses together. First we can consider the Combination Hypothesis, which combines all three. It says that physical space-time and its contents were created by beings outside physical space-time, that microphysical processes are constituted by computational processes, and that our minds are outside physical space-time but interact with it.

As with the hypotheses taken individually, the Combination Hypothesis is coherent, and we cannot conclusively rule it out. And like the hypotheses taken individually, it is not a skeptical hypothesis. Accepting it might lead us to revise a few of our beliefs, but it would leave most of them intact.

Finally, we can consider the Metaphysical Hypothesis (with a capital M). Like the Combination Hypothesis, this combines the Creation Hypothesis, the Computational Hypothesis, and the Mind-Body Hypothesis. It also adds the following more specific claim: the computational processes underlying physical space-time were designed by the creators as a computer simulation of a world.

(It may also be useful to think of the Metaphysical Hypothesis as saying that the computational processes constituting physical space-time are part of a broader domain, and that the creators and my cognitive system are also located within

this domain. This addition is not strictly necessary for what follows, but it matches up with the most common way of thinking about the Matrix Hypothesis.)



The Metaphysical Hypothesis is a slightly more specific version of the Combination Hypothesis, in that in specifies some relations between the various parts of the hypothesis. Again, the Metaphysical Hypothesis is a coherent hypothesis, and we cannot conclusively rule it out. And again, it is not a skeptical hypothesis. Even if we accept it, most of our ordinary beliefs about the external world will be left intact.

IV. The Matrix Hypothesis as a Metaphysical Hypothesis

Recall that the Matrix Hypothesis says: I have (and have always had) a cognitive system that receives its inputs from and sends its outputs to an artificially-designed computer simulation of a world.

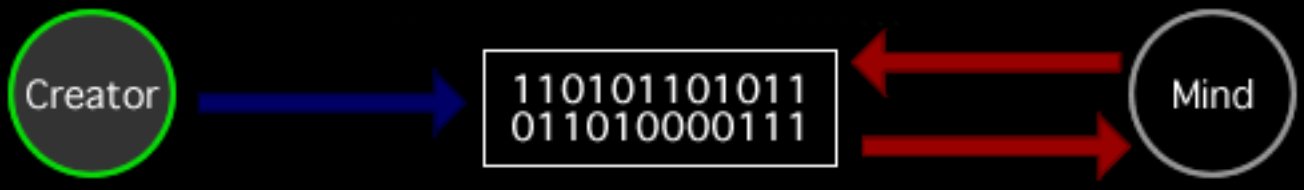
I will argue that the Matrix Hypothesis is equivalent to the Metaphysical Hypothesis, in the following sense: if I accept the Metaphysical Hypothesis, I should accept the Matrix Hypothesis, and if I accept the Matrix Hypothesis, I should accept the Metaphysical Hypothesis. That is, the two hypotheses *imply* each other, where this means that if one accepts the one, one should accept the

other.

Take the first direction first, from the Metaphysical Hypothesis to the Matrix Hypothesis. The Mind-Body Hypothesis implies that I have (and have always had) an isolated cognitive system which receives its inputs from and sends its outputs to processes in physical space-time. In conjunction with the Computational Hypothesis, this implies that my cognitive system receives inputs from and sends outputs to the computational processes that constitute physical space-time. The Creation Hypothesis (along with the rest of the Metaphysical Hypothesis) implies that these processes were artificially designed to simulate a world. It follows that I have (and have always had) an isolated cognitive system that receives its inputs from and sends its outputs to an artificially-designed computer simulation of a world. This is just the Matrix Hypothesis. So the Metaphysical Hypothesis implies the Matrix Hypothesis.

The other direction is closely related. To put it informally: If I accept the Matrix Hypothesis, I accept that what underlies apparent reality is just as the Metaphysical Hypothesis specifies. There is a domain containing my cognitive system, causally interacting with a computer simulation of physical-space time, which was created by other beings in that domain. This is just what has to obtain in order for the Metaphysical Hypothesis to obtain. If one accepts this, one should accept the Creation Hypothesis, the Computational Hypothesis, the Mind-Body Hypothesis, and the relevant relations among these.

This may be a little clearer through a picture. Here is the shape of the world according to the Matrix Hypothesis.



At the fundamental level, this picture of the shape of the world is exactly the same as the picture of the Metaphysical Hypothesis given above. So if one accepts that the world is as it is according to the Matrix Hypothesis, one should accept that it is as it is according to the Metaphysical Hypothesis.

One might make various objections. For example, one might object that the Matrix Hypothesis implies that a computer simulation of physical processes exists, but (unlike the Metaphysical Hypothesis) it does not imply that the physical processes themselves exist. I will discuss this and other objections in later sections. For now, though, I take it that there is a strong case that the Matrix Hypothesis implies the Metaphysical Hypothesis, and vice versa.

V. Life in the Matrix

If this is right, it follows that 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, just 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. When do our nonphysical minds start to exist? It depends on just when new envatted cognitive systems are attached to the simulation (perhaps at the time of conception within the matrix, or perhaps at time of birth?). Is there life after death? It depends on just what happens to the envatted systems once their simulated bodies die. How do mind and body interact? By causal links that are outside physical space and time.

Even if we 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 are not envatted, but are 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.

VI. Objection: Simulation is not Reality

(This slightly technical section can be skipped without too much loss.)

A common line of objection is that a simulation is not the same as reality. The Matrix Hypothesis implies only that a simulation of physical processes exists. By contrast, the Metaphysical Hypothesis implies that physical processes really exist (they are explicitly mentioned in the Computational Hypothesis and elsewhere). If so, then the Matrix Hypothesis cannot imply the Metaphysical Hypothesis. On this view, if I am in a matrix, then physical processes do not really exist.

In response: My argument does not require the general assumption that simulation is the same as reality. The argument works quite differently. But the objection helps us to flesh out the informal argument that the Matrix Hypothesis implies the Metaphysical Hypothesis.

Because the Computational Hypothesis is coherent, it is clearly *possible* that a computational level underlies real physical processes, and it is possible that the computations here are implemented by further processes in turn. So there is *some* sort of computational system that could yield reality here. But here, the objector will hold that not all computational systems are created equal. To say that some computational systems will yield real physical processes in this role is

not to say that they all do. Perhaps some of them are merely simulations. If so, then the Matrix Hypothesis may not yield reality.

To rebut this objection, we can appeal to two principles. First, any abstract computation that could be used to simulate physical space-time is such that it *could* turn out to underlie real physical processes. Second, given an abstract computation that *could* underlie physical processes, the precise way in which it is implemented is irrelevant to whether it *does* underlie physical processes. In particular, the fact that the implementation was designed as a simulation is irrelevant. The conclusion then follows directly.

On the first point: let us think of abstract computations in purely formal terms, abstracting away from their manner of implementation. For an abstract computation to qualify as a simulation of physical reality, it must have computational elements that correspond to every particle in reality (likewise for fields, waves, or whatever is fundamental), dynamically evolving in a way that corresponds to the particle's evolution. But then, it is guaranteed that the computation will have a rich enough causal structure that it *could* in principle underlie physics in our world. Any computation will do, as long as it has enough detail to correspond to the fine details of physical processes.

On the second point: given an abstract computation that could underlie physical reality, it does not matter how the computation is implemented. We can imagine discovering that some computational level underlies the level of atoms and electrons. Once we have discovered this, it is possible that this computational level is implemented by more basic processes. There are many hypotheses about what the underlying processes could be, but none of them is especially privileged, and none of them would lead us to reject the hypothesis that the computational level constitutes physical processes. That is, the Computational Hypothesis is

implementation-independent: as long as we have the right sort of abstract computation, the manner of implementation does not matter.

In particular, it is irrelevant whether or not these implementing processes were artificially created, and it is irrelevant whether they were intended as a simulation. What matters is the intrinsic nature of the processes, not their origin. And what matters about this intrinsic nature is simply that they are arranged in such a way to implement the right sort of computation. If so, the fact that the implementation originated as a simulation is irrelevant to whether it can constitute physical reality.

There is one further constraint on the implementing processes: they must be connected to our experiences in the right sort of way. That is when we have an experience of an object, the processes underlying the simulation of that object must be causally connected in the right sort of way to our experiences. If this is not the case, then there will be no reason to think that these computational processes underlie the physical processes that we perceive. If there is an isolated computer simulation to which nobody is connected in this way, we should say that it is simply a simulation. But an appropriate hook-up to our perceptual experiences is built into the Matrix Hypothesis, on the most natural understanding of that hypothesis. So the Matrix Hypothesis has no problems here.

Overall, then, we have seen that a computational process *could* underlie physical reality, that any abstract computation that qualifies as a simulation of physical reality could play this role, and that any implementation of this computation could constitute physical reality, as long as it is hooked up to our experiences in the relevant way. The Matrix Hypothesis guarantees that we have an abstract computation of the right sort, and it guarantees that it is hooked up to our

experiences in the relevant way. So the Matrix Hypothesis implies that the Computational Hypothesis is correct, and that the computer simulation constitutes genuine physical processes.

VII. 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 has virtual hair, so he is correct.

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 exactly what one might 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.

VIII. Other Skeptical Hypotheses

The Matrix Hypothesis is one example of a traditional "skeptical" hypothesis, but it is not the only example. Other skeptical hypotheses are not quite as straightforward as the Matrix Hypothesis. Still, I think that for many of them, a

similar line of reasoning applies. In particular, one can argue that most of these are not global skeptical hypotheses: that is, their truth would not undercut all of our empirical beliefs about the physical world. At worst, most of them are *partial* skeptical hypotheses, undercutting some of our empirical beliefs, but leaving many of these beliefs intact.

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.

The 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 miniscule probability. If I am chaotically envatted, do physical processes obtain in the external world? 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!

David Chalmers

David Chalmers' website: www.consc.net

(Some philosophical notes on this article can be found here.)

Artificial Ethics

Julia Driver

The significance of *The Matrix* as a movie with deep philosophical overtones is well recognized. Whenever the movie is discussed in philosophy classes, comparisons are made with Descartes' *Meditations*, particularly the dream argument and the evil genius scenario, both of which are intended to generate skeptical doubt. How do we know, for example, that we are awake now, rather than merely dreaming? How do we know that our thoughts are not being manipulated, and that our perceptions of 'reality' are accurate? *The Matrix* makes these doubts stand out vividly.

However, *The Matrix* raises many other interesting philosophical issues, and ones that are worthy of further discussion. This essay explores some of the moral issues raised in *The Matrix*. The first is the issue of the moral status of the created beings, the 'artificial' intelligences, which figure into the universe of *The Matrix*. The second is the issue of whether or not one can do anything wrong in circumstances where one's experiences are non-veridical; that is, where one's experiences fail to reflect reality.

I. The Moral Status of Programs

There is a reality to the Matrix. The substance of that reality may differ dramatically from the substance we label 'real' — the 'real' world is the desert reality that Morpheus reveals to Neo. But it is clear that, out of the grip of the Matrix, though still having certain dream-like experiences, Neo and his enlightened friends are dealing with actual sentient programs, and making

decisions that have actual effects for themselves as well as the machines and the programs. What is the moral status of the sentient programs that populate the Matrix, or, for that matter, the moral status of the machines themselves? The universe of *The Matrix* is populated with beings that have been created — created by programmers or created by the machine universe itself. The agents, such as Smith, Neo's pursuer, are prime examples. These beings come into and go out of existence without comment on the part of whoever controls the switches — and without any moral debate on the part of the humans who also would like to see the agents destroyed. There seems to be an implicit view that their existence is less significant, their lives of less moral import, than the lives of 'naturally' existing creatures such as ourselves. An obvious explanation for this attitude is that humans are long accustomed to thinking of themselves as being at the center of the universe. The geographic point changed with Copernicus. However, our view of our dominant place in the moral universe has stayed fixed. But, once again, science — and particularly, now, cognitive science holds the potential for challenging this certainty. And science fiction such as *The Matrix*, which explores differing directions for these potentialities, also brings challenges to this worldview. What *The Matrix* offers is a vivid thought experiment. It is a thought experiment which makes us ask the sort of 'what if?' question that leads to a change in self conception. It forces us to see where our well accepted moral principles would take us within one possible world.

We know that killing human beings is wrong. It is wrong because human beings have moral standing. Human beings are widely believed to have this standing in virtue of consciousness and sentience. For example, a rock has no moral standing whatsoever. Kicking a rock does not harm it, and no moral rights are violated. It is an inanimate, non-conscious object incapable of either

thought or sensation. Animals, however, are generally taken to have some moral standing in virtue of their sentience. Kicking an animal for no compelling reason is generally taken to be immoral. Human beings have greater standing in virtue of their higher rational capacities. They can experience more varied and complex harms, and a wider range of emotional responses – such as resentment – in virtue of their rationality. *How* one came into existence is not taken to be morally significant. Some people are the products of natural conception, and some are the result of conception in the laboratory. This makes no difference to the possession of those qualities we take to be morally significant – consciousness and rationality. And, surely, the *substance* from which someone is created is completely irrelevant to the issue of moral status. If a person's consciousness could somehow be transferred to a metallic or plastic robotic body, the end result would still be a person.

It would seem, then, that the fact that one is created, or artificial, is in no way relevant to one's moral standing. And, if this is the case, then the world of *The Matrix* presents underappreciated moral complexities. Agents such as Smith, while not very pleasant, would arguably have moral standing, moral rights. Of course, Neo has the right to defend himself — Smith is not, after all, an innocent. Indeed, if the religious theme is pursued, he is an agent of darkness. But any innocent creations of the machines — beings brought into existence to populate the Matrix — also would have moral rights. Just as it would be wrong to flip a switch and kill an innocent human being, no matter how that human being came into existence, it would be wrong to flip a switch and kill a sentient program. As long, of course, as that program possessed the qualities we regard as *morally relevant*. And this is where one of the primary issues raised by the possibility of artificial intelligence becomes important to the question at

hand. Do these programs possess consciousness? Since we are considering the world of *The Matrix*, let's look at what evidence seems to exist in the movie. While we don't have much information about the machines themselves, their agents are on ample display.[1]

Smith, of course, and his colleagues seem remarkably without affect. Yet, at critical points they do display emotions: anger, fear, and surprise. They seem able to plan and to carry through on a plan. Smith also displays a capacity for sadistic pleasure — at one point he displays this, when he forces Neo's mouth shut. Smith also displays extreme fear near the end of the movie, when Neo leaps through him. The agents display many, if not all, of the responses we *associate* with consciousness and sentience. But this brings us to another skeptical challenge posed in *The Matrix*. How can we be sure they do posses minds, and are not mere automata, albeit highly complex ones? Though the movie invites this reflection, it is important to see where this challenge can take us. The "how can I be sure?" question can extend beyond the agents to our fellow human beings. Since a person's conscious experiences are essentially private, one cannot be directly aware of another's experiences. We might try, as St. Augustine suggested, to solve this problem by appeal to analogy: I do directly experience my own mental states — I know that I am a conscious, aware, being. I also know on the basis of observation that I am structurally similar to other human beings. Thus, I reason by analogy, that they must experience mental states as well.[2] And, indeed, The Matrix invites such a comparison when the agents display behavior consistent with the experience of certain psychological states.[3]

Given, then, that we believe what we are invited to believe it would follow that the sentient programs, the cyber persons, do possess those qualities we

associate with moral standing. They have moral rights on the basis of consciousness and sentience and rationality. Thus, their moral standing is the same as that of human beings.

It is possible that human beings have some additional value — a kind of antiquarian value. We are, so to speak, "the originals." The original Mona Lisa, for example, has value in excess of its copies. But this kind of value is not moral value and does not reflect on the moral standing of the object, or the moral significance of the lives themselves. The Mona Lisa does have value, but no moral standing since it is a mere painting; it lacks consciousness. It may be damaged, but not harmed in the way that humans and sentient creatures can be harmed.

Perhaps the machines view humans this way. To the machines, the value of humans is mainly instrumental. They are valued as a source of energy, but they may also have some antiquarian value. Humans are merely relics of a past they themselves helped to destroy. If that's the case, the machines have turned the tables. They are making the same moral mistake humans apparently made in the context of The Matrix, in viewing other rational life forms as simple instruments, to use and destroy as one wishes. Indeed, both sides of the conflict seem to have displayed some moral blindness. The humans, in using and destroying, and the machines, certainly, in their subjection of the humans. But both sides view themselves as fighting for survival, and I imagine that Smith and Smith's creators, as well as Neo and his friends, would argue that moral qualms like these are a luxury.

II. Manipulation and Immorality

The world that the pre-enlightened Neo inhabits is one made up by machines.

The machines have created a humdrum existence for humans, to keep them happy and pacified and free of the knowledge that they are being used as a source of energy for the machines. Most humans believe that this world is real, but they are mistaken. Within this world they build lives for themselves, have relationships, eat lovely dinners, and at least seem to both create and destroy. To some extent this existence is dream like. It isn't real. When the unenlightened person thinks he's eating a steak, he isn't. Instead, the machines generate mental experiences which correspond to the experience of eating a steak, but which are non-veridical – that is, the person is not *actually* eating a steak. There is no real or actual steak. The human being's actions, in that respect, have no real or actual consequences in a world that exists independently of his or her mind. However, even in this unenlightened state, the humans do have *some* control, since what they 'do' in the Matrix has consequences which are realized in the real world. Getting smashed by a truck in the Matrix kills the person in reality. The Matrix offers a 'brain-in-a-vat' experience, but one where the experiencer does have some control.[4] The enlightened can, in principle, understand the rules of the Matrix and learn to exert that control with full understanding.[5]

But, as the steak example illustrates, there are many other 'actions' they perform that seem to have no effects in the real world. The pre-enlightened Neo and most of the humans living in the Matrix seem to be deluded. One issue raised by this is the extent to which they can be held responsible for their actions in the Matrix. Suppose, just for the sake of argument, that something like wearing fur is immoral. Is simply making a choice to wear fur, along with the belief that one is wearing fur, enough to make one guilty of wrongdoing? Is it really *only* the thought that counts, morally? A competing view is that the

choices people make must result in actual bad consequences in order for them to be guilty of wrongdoing; or, actual good consequences in order for them to be considered to have acted rightly. So, the issue is that of whether or not the moral quality of a person's actions — its rightness or wrongness — is determined solely by his or her subjective states, or whether, instead, actual consequences figure into this determination.

In the Matrix if fur is worn it is virtual fur, and not real — though the wearer does not realize this. Again, this is because he or she is being mentally manipulated. But is this a genuine delusion? Certainly, an insane person who fails to have a grip on reality, and is deluded in this sense, is thought to have *diminished* moral responsibility for what he or she does while deluded. Such a person is generally held to not be morally responsible in those circumstances. He is not punished, though he may be confined to a mental hospital and treated for his insanity. The explanation is that the actions performed while insane are not truly voluntary. If the persons who live in the Matrix are similarly deluded, then it would seem that they are not responsible for what they 'do' in the Matrix.

Some writers have argued that one cannot be held responsible for what happens in a *dream*, since dreams themselves are not voluntary, nor are the 'actions' one seems to perform in a dream.[6] Other writers, such as Henry David Thoreau, had the view that what we seemed to do in a dream reflected on our character; and the contents of dreams could reveal true virtue or vice.[7] Even if the actions one performs in a dream have no actual good or bad consequences, they reveal truths about one's emotional make-up, and one's inner desires, and these, in turn are revealing of character. But, as we've discussed, the Matrix isn't a dream. The unenlightened exist, rather, in a state

of psychological manipulation. The actions they seem to perform don't always have the effects (in reality) that they have reason to expect, based on their manipulated experiences. But even in the Matrix we can argue that they make voluntary choices. They are not irrational. They are not like the insane. Neo believes what any rational, reasonable person would believe under the circumstances. The pre-enlightened are analogous to persons who make decisions based on lies that others have told them. They act, but without relevant information. It's that condition that Neo would like to rectify at the end of *The Matrix*.

The view I favor is that without *actual* bad effects the actions of those in the Matrix are not immoral. But, again, this claim is controversial. Some would argue that it's simply "the thought that counts"; that it is the person's intentions which determine the moral quality of what he or she does. Immanuel Kant, for example, is famous for having claimed that all that matters, intrinsically, is a good *will* – actual consequences are irrelevant to moral worth.[8] However, it would then be the case that forming bad intentions in one's dreams is also sufficient for immorality, and this seems highly counterintuitive. If that's true, then the intention to do something immoral along with the belief that one has so acted, is enough to make one guilty of moral wrongdoing. Instead, it seems more plausible that it must also be the case that there is some actual bad brought about, or at least the realistic prospect of some actual bad consequences, and thus non-veridical 'wrongdoing' in the Matrix is not actual wrongdoing.

This seems to be clearly the case in a dream. In a dream, when the dreamer decides to do something bad that decision doesn't impact on the real world. But the Matrix is not really a dream. If we assume that the virtual world of the

Matrix is *complete* — that is, completely like the real world before the machines took over — then the virtual 'harms' are still real in that they are realizied in terms of *actual* unpleasant mental states. The virtual fur coat is the result then of a virtual animal getting killed, but a virtual animal with all the right sorts of mental states — in this case, pain and suffering. If this is the case, then the killer, though mistaken in thinking the dead animal 'real' has still produced bad effects in the form of genuine pain and suffering. And thus, the action is immoral even though non-veridical. However, if the world of the Matrix is incomplete, the issue becomes more complicated. If Cypher's virtual steak comes from a virtual meat locker, and the meat locker is the end of the line — and the acquisition of the steak does not involve the killing of a virtual animal with all the same psychology of pain and suffering a 'real' animal feels, then no moral harm has been done.

But note that Thoreau's point still holds even though the Matrix is not exactly like a dream. That is — even if a person hasn't actually done anything bad, or caused any real harm to another sentient life form, we may still make a negative evaluation of the person's *character*.

But my guess is that the Matrix is a complete alternate reality created in the image of the pre-machine reality. And the Matrix, if it does offer such a complete replication of the pre-machine reality, is truly a self-contained world. It has its own objects, its own people, animals and ... ethics. The systematic deception of the humans doesn't change this.

Julia Driver

Footnotes

1. The issue of the moral status of the machines themselves should be kept distinct from the issue of the moral status of the sentient programs. I will focus on the latter issue here in discussion, simply because the movie provides more information about the behavior of these constructs. But the same points would hold for the machines themselves – if they have those qualities that are morally significant, consciousness and rationality, then they also possess moral standing.

2. St. Augustine, *The Trinity* (8.6.9). Again, this line of reasoning is controversial since it relies on a single case analogy.

3. A lot hinges on what we take to be 'structurally similar'. Some would argue that while the sentient programs are not themselves structures, the machines are, and thus the machines may possess consciousness, though the programs cannot. However, I believe the sentient programs can be structurally similar if that's understood functionally – their code has structure which provides functional equivalence to the physical states that underlie our mental states. But, this issue would be extremely controversial, and there isn't enough time to delve into it more fully here.

4. See Christopher Grau's introductory essays on this site for more on dream skepticism and brain-in-a-vat skepticism.

5. The unenlightened, on the other hand, are constantly being "Gettiered". A woman may have justified true belief that her husband is dead, because she has just 'seen' him smashed by a truck. But being in the Matrix she lacks true knowledge because she is deceived in the true manner of his death.

6. See, for example, William Mann's "Dreams of Immorality," *Philosophy* (1983), pp. 378-85.

7. Thoreau writes about this in *A Week on the Concord and the Merrimack* (1849).

8. This also is controversial, but see Kant's *Foundations of the Metaphysics of Morals*, trans. Lewis White Beck, and critical essays ed. by Robert Paul Wolf (NY: MacMillan, 1969):

> Nothing in the world — indeed, nothing even beyond the...world — can possibly be conceived which could be called good without qualification except a good will...The good will is not good because of what it effects or accomplishes or because of its adequacy to achieve some proposed end; it is good only because of its willing, i.e., it is good of itself. (pp. 11-12)

Neo's Freedom... Whoa!

Michael McKenna

The Matrix provides a fine resource for illustrating philosophical ideas. Many films have themes that one can philosophize about, or that serve as useful illustrations of philosophical ideas, such as the wonderful films *Sophie's Choice* or *The Sheltering Sky*. But *The Matrix* offers more than this. It belongs in a special class of films including *Blade Runner*, *Total Recall*, *Crimes and Misdemeanors*, *A Clockwork Orange*, *The Unbearable Lightness of Being*, and *The Truman Show*. All of these films are intentionally philosophical. Each shows how richly philosophical themes can be developed through cinema. Perhaps the best of these films is *The Matrix*.

I.

No doubt, the most striking philosophical theme found in *The Matrix* concerns skepticism about knowledge of an external world. The dream world Neo inhabited was a perfectly comfortable "reality"—except for the fact that it was not reality.[1] Life from inside it completely shielded one from what Morpheus aptly called "the desert of the real," that desolated shell of a planet on which countless humans were unknowingly ensconced in slimy wombs. But there are many other philosophical themes explored within *The Matrix*. One is the concept of freedom. Freedom is mentioned at various points in the film.[2] It mattered a great deal who did what freely. For instance, it was important that Neo freely chose to take the red pill and not the blue pill. Had he taken the blue pill, he'd have been returned to that humdrum dream world of vapid city dwellers. He'd never have taken the path that eventually led him to his heroic

defeat of the agents, and that left him at the end of the film entertaining the prospect of saving the human race. At various other points Neo made choices freely, and, as with taking the red pill, it was the quality of having made them freely that gave them the importance they had. For instance, Neo freely decided to risk his life for Morpheus; instead of fleeing when his own life was in danger, he returned to save Morpheus from cranial meltdown at the hands of those treacherous agents in their zoot suits. Also, Neo freely followed the white rabbit that led him tumbling down that rabbit hole. And he remained in the car when Trinity and Switch gave him the opportunity to bail. By remaining in the car, Neo freely chose to resist the agents. He chose on his own not to get out and walk away down that street, down that well worn path that, Trinity reminded him, led to nowhere special. In choosing to remain in the car, he freely embarked upon a path that would lead to an exciting future, to an exciting life.

But it was not just Neo's freedom that mattered. Freedom was an issue for the others as well. During Cypher's attempted mutiny, Trinity reminded him that all of Morpheus's rebels had freely chosen the red pill, and so none could claim that they were in their dire straights undeservedly. All the same, Cypher regretted his choice. He felt duped; he did not regard his choice to take the red pill as free. As he saw it, he was scammed. In fact, he was of the opinion that he'd have had more freedom as a steak-eating, satiated participant in *The Matrix*, oblivious to the "truth" about the ugly shell that would have held him in perpetual slumber.

Freedom also mattered a great deal when it was *not* possessed. It seems that this was the case with those countless human drones, all contained in their artificial wombs. As Morpheus and company saw it (save for Cypher), their

poor, ignorant kin were victims, blind to their lack of freedom—maybe even happy in their plodding little lives within the Matrix, working in cubicles all day—but victims all the same, enslaved in the service of generating battery juice for those battery-powered A.I. meanies. Even the leader of the agents' posse, Agent Smith, valued freedom. He too was limited in his freedom since he was required to do something against his will, namely remain in the Matrix and deal with those pesky rebel infiltrators. As he confessed to Morpheus, he hated having to be there, hated the smell of the humans. He felt trapped. Poor guy. In the end, Agent Smith's freedom was dramatically impaired by a liberated Neo, who had turned the tables and was now screwing with him.

But of course, all of this is to leave the concept of freedom unanalyzed, and to take the claims of freedom within the film on face value. As any good student of philosophy is aware, there are quite general skeptical challenges to (certain kinds of) freedom that might undermine the very idea that any agent is free in at least one important respect. Let's defer for just a bit longer placing any theoretical structure on what freedom might be, and on the sorts of challenges there might be to it. Let's fix upon some further observations that will subsequently help us to bring into clear focus a few frequently unacknowledged but powerful points about the freedom of human agency, a freedom many have called freedom of the will.

It appeared in the film that some had more freedom than others. Morpheus's crew was amazed watching Neo fight Morpheus for the first time. They thought that the untrained neophyte Neo was just so fast, faster than any of the others. Their hope was that Neo was "The One". No doubt there are biblical themes throughout the film, and no doubt "The One" is one of those themes; "The One" is something like a divine savior. A crucial feature of this savior is

that whoever could fill the bill would have more freedom within the Matrix than could any other rebel visitor to it, or for that matter, any other intentional being operating within the Matrix, including the agents. Indeed, their hope was that Neo's freedom within the Matrix would be like that of God; Neo would have unlimited freedom. So it appeared that Neo, even when first getting acquainted with his abilities, had more freedom within the Matrix than did Trinity, Cypher, or any of the rest of Morpheus's gang (save for Morpheus himself). But there are other comparisons as well that indicate different degrees of freedom within the Matrix. Neo, Morpheus, and all of the rebels had more freedom within the Matrix than did all those clueless characters walking the streets, living in their homes, watching the TV, going to work, etc. At least as Morpheus and company saw it, the clueless were completely unfree.

Until near the film's end, Neo had *less* freedom than did the agents. The agents could simply move about satisfying most any desire they had, taking on others' bodies, appearing whenever and wherever they wanted, and operating with fantastic foresight about who would be where, when, etc. These agents defied what seemed to be the laws of nature (as structured within the Matrix). They could emerge unscathed after being slammed by speeding trains that would have crushed and destroyed any run of the mill putz living out his ordinary life within the Matrix. They took bullets and kept a tickin', and they could simply make a person's mouth disappear at will. They had the run of the place, at least until those closing moments of the film. But in those closing moments of the film, Neo was the freest agent operating within the Matrix. Hell, by the time he came to realize his true potential within it, he could beat the crap out of those battery-powered robot-demons, stop bullets, and fly... like Superman.

One more very important observation before we roll up our sleeves and do some philosophical work: The special sort of freedom that Neo seemed to possess in the film was a freedom confined to the Matrix. The same, of course, applies to Morpheus and the other rebels whom Morpheus trained. The film has given us no reason to believe that Neo, or anyone else, has any special freedom outside the Matrix. In the "real" world, as it is in the space ship with those nasty flying bugs out hunting down rebel ships on that desolated planet, Morpheus, Neo, Trinity, Cypher, and the rest of the clan are just normally functioning human agents like you or me. Presumably, in the real world, Neo's just a guy, a guy who, analogous to poor, impaired, nobody Tommy in The Who's rock opera *Tommy*, is transformed in game mode to the most gifted being ever to play the relevant game—a pinball wizard. In the Matrix, that is, roughly, in the ultimate of video game consoles, Neo ain't got no distractions, can't hear no buzzes or bells, always gets the replay and never tilts at all.

So *in The Matrix*, near the end of the film, as Neo comes to master the game, he's totally dialed in. It's gotta rock! Let us call this freedom that Neo possesses within the Matrix *absolute freedom*, and let us call the feature that seems to go with it *the property of rocking*. No doubt, when Neo first saw such amazing freedom exercised—when Morpheus leapt an incredible distance from one skyscraper to another—he judged that indeed such extreme freedom did rock, and in amazement he appropriately expressed himself thusly: "Whoa!"

II.

The concept of absolute freedom and its presumed property of rocking will be further developed in the closing sections of this essay. But for now, let us first give some theoretical structure to the idea of freedom, forgetting about

absolute freedom, and let us consider briefly a classical philosophical challenge to it. Once we have these issues in place, we'll turn back to the film and examine our natural reactions to it, reactions such as the many mentioned above.

The term *freedom* is used in many contexts, and there is no reason to assume that there is a single meaning of the term. Minimally, all of the uses of the term do seem to share the feature that resistance of some sort, encumbering or impeding desired conduct, gets in the way of freedom. Typically, one is not free when she is frustrated in some manner from unencumbered pursuit of her desired course of action. But the absence of impediments is clearly not sufficient for the kind of freedom that mattered to Morpheus, Neo, and company, nor to what is valuable and distinctive of the human condition. A stupid dog can sometimes act unencumbered when, for instance, she is unleashed—when she is set free. And though free in a very basic way, the stupid dog's freedom is not the kind that makes philosophers, theologians, politicians, moralists, or just your run of the mill high-minded folk get the warm fuzzies. No. The freedom worth talking about seems to be a freedom distinctive of persons, and this suggests that understanding the relevant notion of freedom first requires an understanding of what it is to be a person.

Regrettably, offering an account of personhood is beyond the scope of this essay. But to appreciate what seems to mark persons from non-persons, those familiar with the movie *Blade Runner* can reflect upon the characters Decker and the replicant Rachael, with whom Decker fell in love. Although Decker was a human being (maybe), and Rachael was an artificial replicant of a human being, both were *persons*.[3] Both were capable of planning lives, of developing intimate relationships of love and hate, of fearing for, and finding dear, their

own lives, and the lives of other persons. Both had the capacity for abstract thought, emotional responses to others, self-consciousness, etc. Less developed cognitive creatures were not persons, such as the primitive little A.I. machines that kept J.F. Sebastian company (J.F. Sebastian was another character in *Blade Runner*). Or to draw upon other clear illustrations of personhood from other sources in film, E.T. from the classic Spielberg movie was a person. Data from the *Star Trek* series and movies is a person, though neither E.T. nor Data is a human being. So, for our purposes, Neo, Morpheus, Trinity, as well as the agents, are all persons—though the agents, like E.T. or Data, are non-human persons.

Even restricting the term freedom to its applications to persons, there are at least two sorts that have been the focus of a great deal of philosophical attention for well over two millennia now. One is a matter of political freedom, another is a matter of metaphysical freedom, the latter being understood as freedom of the will. Political freedom concerns the freedom of persons to conduct themselves as they see fit within the political landscape. The nature of the political landscape is itself a matter of dispute. Does the landscape germane to political freedom include economic empowerment? Or does it merely involve what are often referred to as the civil liberties, such as the liberty to speak unthreatened from harm of prohibition, to organize as one wishes, etc? Political freedom, whatever it comes to, is certainly a deeply important sort of freedom, and no doubt, it is a sort of freedom that Morpheus was struggling to give back to the human race. At least this is how Morpheus and his comrades saw it. But the more immediate sort of freedom to which the film directs our attention is not political freedom, but metaphysical freedom, that is, freedom of the will.

Before turning our attention to the topic of *free* will, it is worth asking, *what is a will?* This is also the subject of a great deal of dispute, but it is natural to think of the will as the aspect of a creature's mentality that is the source of voluntary, intentional (that is, goal-directed) action. Hence, any agent—that is, any being that acts, such as a dog, a cat, a chimpanzee—has a will. The philosophical gem worthy of reflection is what makes a will free, and most notably free in the special way distinctive of a unique class of agents, those who are persons.

A word of caution: The expression "metaphysical freedom" is often regarded derisively by theorists, largely outside of philosophy, who fallaciously associate it only with extravagant views about the human condition, such as the view that metaphysical freedom provides persons with a capacity to transcend the material world, to choose and act unlimited by the laws of nature, or by any constraints from the material world. And while some theories of free will do attribute to persons the ability to perform 'very small' miracles whenever they act freely[4], all the expression *metaphysical freedom* need pick out is a distinctive feature of personhood—a feature unique to the *will* of a person, perhaps part of the essence or the nature of what it is to be a person. How to understand this freedom is up for grabs. So, to be clear: the very mention of the notion of metaphysical freedom, or freedom of the will, does not *entail* anything mysterious. It does not entail anything contrary to the spirit of an inquiry such as Darwin's, or that of the neurobiologist. It might turn out that free will involves no special miraculous features of agency at all, that metaphysical freedom is entirely consistent with a deflationary account of human persons according to which all human persons are entirely the products of their genetics, their environment, and any other physical factors impinging upon them. That said, it should be kept in mind that, on the other hand,

serious philosophical reflection might indicate that the concept of free will implies that a deflationary view of persons is false. But the crucial point here is that it is not part of the meaning of the very term *metaphysical freedom*, or *freedom of will*, that it involve anything spooky, mysterious, unworldly, or otherwise beyond the pale of what is in principle explicable in terms of our best natural sciences.

III.

Here is a theory-neutral characterization of free will:

> Free will is the ability of persons to control the future through their choices and actions.

This is a lean definition that is not biased towards any one particular manner of philosophizing about free will. Of course, it is only a first pass and cries out for refinement. The crux of the issue concerns how best to articulate the ability to control the future. Let us consider two ways to articulate further this characterization of free will.

It is quite natural to assume, as many philosophers do, that a person acts with freedom of the will only if there are alternative courses of action available to her at the time at which she acts. On a model such as this, a person's freedom of the will consists partly in her being in control of a spectrum of options that, so to speak, open up different temporal paths, allowing her access to different unfolding futures, different ways that her life might go. At various points in the film, this picture of freedom was emphasized, as when Neo chose to remain in the car and not bail when Trinity and Switch gave him the opportunity to do so. This picture of freedom was also highlighted when Neo chose to return and fight the agents so as to save Morpheus. Instead Neo could have left Morpheus

to (what seemed to be) his inevitable demise.

So one way to advance free will is in terms of *alternative possibilities*. But there are other strategies for understanding free will, strategies that might or might not work in tandem with a demand for alternative possibilities. For instance, another way to think about free will is in terms of what *does* happen, what an agent *does do*, and not in terms of what other things she might do or might have done. Instead of focusing on alternative possibilities, this manner of theorizing concentrates upon the *source* of an agent's actions. On this approach, freely willed actions arise from certain salient features of an agent's self, features that indicate that, in an important respect she—*the agent*—is the source of how the future does unfold. To illustrate, consider a paradigmatic case of an agent who lacks free will. An unwilling addict, for example would not act with freedom of will when she takes the drug to which she is addicted. This is because her addictive desire to take the drug is so strong that it compels her to take it even though she is unwilling in taking the drug. *She* does not desire that her desire for the drug cause her to take it. But she does take it all the same. The future does not unfold as she herself would like it to unfold. On the other hand, sometimes properly functioning persons *do* act precisely as they wish (however "as they wish" might be understood). When they do, if all goes well, the future unfolds as they would like it to unfold, and it unfolds in this way partially *because what they do causes it to unfold in this way*. Hence, in a very basic way, these normally functioning persons are guiding how the future unfolds when they act unencumbered. They are the ones bringing about certain events, shaping the future in certain ways via their agency. *They are sources of control over the future*. It should also be clear that Morpheus and Neo illustrated such views of freedom. They certainly were at points sources of

"control" over how their futures were unfolding. Morpheus and Neo, as well as the rest of the rebels, were making *their* marks inside and outside of the Matrix. Much to the chagrin of the agents, Morpheus and his crew were sources of control over how certain events were unfolding.

In summary, if we understand free will as a capacity of persons to control the future through their choices and actions, then there are two ways that one might further develop this idea of control over the future. One is in terms of control over alternative possibilities; another is in terms of one's very self being a source of how the future goes, an authentic shaper or causer of events in the world.

IV.

However the concept of free will is developed, there is a classical challenge to the very idea that any person possesses it. In particular, some philosophers believe that if the universe is fully determined, then no person has free will. What it means to suggest that the universe is determined is a distinct and controversial philosophical topic. A currently fashionable definition of *determinism* has it, roughly, that the past, combined with the laws of nature, causally insures one unique future. To appreciate fully this definition, one needs an account of what the past is (or the facts of it), what it means to causally insure, etc. But the general idea is basically captured with the suggestion that, for any person, states of the world independent of that person, or independent of features of her intentional agency (possibly, states of the world prior to her birth), combined with the laws governing the natural world (such as the laws of physics, chemistry, biology, etc.), are themselves sufficient to fix fully what that person does at any time. Crudely put, are

persons and their conduct exhaustively explained in terms of their hereditary, their biology, such as their neurobiological functioning, and the environmental influences impinging upon them? Put even more crudely, is all human conduct purely a matter of nature and nurture? Or is determinism false, and is it instead the case that these influences do not all by themselves explain exactly what a person does at any time? If not, does the person herself contribute something over and above these other factors that accounts for why she does what she does?

Incompatibilists believe that if determinism is true, no one has free will. No one can control her future since the universe, so to speak, is really controlling it, and persons and their conduct are merely conduits through which the forces of nature operate. The universe leads some people to act in certain ways, and others to act differently. Persons are not at the helms of their lives, guiding their futures. Persons are products of the universe, not agents freely acting upon it!

Turning to the two ways of developing the concept of free will suggested above, the incompatibilists will argue that either way conflicts with the assumptions of a deterministic world. Suppose that the concept of free will is developed in terms of alternative possibilities. If determinism is true, and if facts distinct from a person's intentional agency, combined with the laws of nature, entail that an agent's intentional conduct will be thus and so, then an agent is not free to do other than thus and so. She has no alternatives over which to exercise control. Her past and the forces of nature have settled for her what path into the future she will take.

Or suppose instead that the concept of free will is developed in terms of an agent's being an *actual source* of how the world goes, and it going that way, at

least in part, *because of her*. If determinism is true, then there are facts prior to any person's birth, combined with the laws of nature, that provide sufficient conditions for how the future will unfold. A person's agency, given determinism, seems to be nothing but a conduit, a facilitator, for what has already been set in motion. She, *ultimately*, is not the source of her action, the controller of an unfolding future. Sure, sometimes the future unfolds as she desires that it does, and sometimes her desires figure in the causes that explain why it does unfold as such. But these very desires, her beliefs, value judgments, her preferences about what motivational states are the ones that she wishes to act upon, *all of these factors* are themselves not factors ultimately issuing from her, but from the determined universe and the unfolding future that is an upshot of it.

As initially puzzling as it seems, *compatibilists* maintain that persons can have free will even if determinism is true. Some compatibilists, embracing a view of free will that requires alternative possibilities, have attempted to show that a determined person might still, in some meaningful sense, have the ability to do other than what she does. Other compatibilists have instead emphasized how an agent might, via her own motivational states, still count as a significant actual source of efficacy in the way the future comes about.[5]

V.

There are various ways in which the tension between compatibilism and incompatibilism is brought out in the film. One is in terms of reflections upon fate. Another is in terms of the Oracle's ability to know the future. Yet another has to do with the status of those poor "enslaved" humans.

It is worth noting that within the film, as in ordinary discourse, the term *fate* is used in two different sorts of ways, ways that are easy to confuse, but upon reflection are clearly distinct. Sometimes *fate* is used to mean what is also meant by *determinism*. This certainly seems to be the primary manner in which it is used within the film. Given this usage, what it is for something to be fated is for it to be causally insured by prior conditions. This view is entirely consistent with one's conduct being a crucial factor in what is causally insured. But on a different construal, if some outcome is fated, then it will come about *no matter what one does*. On this view, one's agency is an idle factor. A certain future will transpire *irrespective of anything one might do*. The standard example of this is the story of Oedipus. The gods were going to see to it that Oedipus met his terrible fate—killing his father and copulating with his mother—no matter what different things were done by any mortal to avoid that outcome.

These two notions are extremely different. To illustrate: If it was fated irrespective of what anyone did that Kennedy would be assassinated on the day he was, then no matter what Lee Harvey Oswald did (including not assassinate anyone), Kennedy was going to be assassinated (by someone). But if it was fated just in the sense of being determined that Kennedy was going to be assassinated, then it mattered a great deal precisely what Oswald did. Had he not done what he did, then Kennedy would not have been shot. One account of fate states that a certain future will unfold *no matter what any person does or will do;* another sates that a certain future will unfold *precisely because of what does or will take place* (which includes, among other things, what people actually do). Typically, philosophers reserve the term *fatalism* for

the former notion and *determinism* for the latter. But for purposes of analyzing the film, let us distinguish between *no-matter-what-one-does fatalism* and *deterministic fatalism*.[6]

When Neo and Morpheus first met, Morpheus asked Neo if he believed in fate. Neo said that he did not since he did not like the idea that he did not control his life. Note that, at this point in the film, what Morpheus meant by fate, and what Neo took it to mean, remained ambiguous between the two notions distinguished above. This is because, if one's life is subject to no-matter-what-one-does fate, then that would undermine one's control with respect to the fated outcome. So Neo's reply could have been in response to the suggestion that life was no-matter-what-one-does fated.[7] Perhaps what Neo found objectionable about fatalism was the thought that his agency in the world would have no effect on the world's outcome at all—no matter what he did. And indeed, that is how it seemed the enslaved humans lived within the Matrix, having no effect no matter what they did on their contribution to generating electricity for the A.I. meanies. But even if this is what Neo meant in that first conversation with Morpheus, later in the film it is clear that Neo also wanted to resist deterministic fatalism. He was committed to the idea that deterministic fatalism would undermine his control over the world. At points it was quite clear that his worry was in the form of alternative possibilities. He resisted the idea that the Oracle could know which of the possible futures before him would be his inevitable actual future. He thought that it was up to him what that future would be—would he choose to save Morpheus or himself? But Neo also seemed to think in terms of source models of control: As he saw it, it was not settled in advance how he would act; he would be the settler of it! As the Oracle was bidding Neo farewell, she herself put those words in his mouth.

Neo, it seems, was an incompatibilist.

If Neo is the incompatibilist in the film, Morpheus is certainly the compatibilist. He believed in his consultations with the Oracle that the future was deterministically fated, that The One would come. But he also believed that what he did, and what the others did, mattered very much to that outcome. (So he certainly did not endorse no-matter-what-one-does fatalism.) Even more importantly, he believed that it mattered very much that what people did, they did of their own free will, hence the use of the blue and the red pills. His advice to Neo was especially telling. Thinking in terms of source control, Morpheus explained to Neo that it is not enough to know that you are The One, you have to *be* The One. That is, Neo had to be the actual source of that special person, which was a matter of his actual conduct in the world, and not merely something he conceptually grasped.

And what of the Oracle herself? To correct the impression that perhaps the Oracle is not really able to foresee the future, Morpheus tells Neo that the Oracle never intended to speak truthfully to Neo about what she foresaw. She only intended to say to Neo what he needed to hear (which of course she knew since she was an Oracle). Surely, if she did make any judgments about what Neo needed to hear, then she did believe that what he would do would matter to how the future would go. If so, then like Neo and Morpheus, she also did not believe in no-matter-what-one-does fate. But being an Oracle, she probably at least entertained the idea that deterministic-fatalism was true. Suppose she did believe it. Was she a compatibilist or an incompatibilist? Might she have believed, consistent with incompatibilism, that all the human struggles to shape the future were unfree actions set in motion by a long, deterministically fated history? Or did she instead, consistent with compatibilism, foresee and understand Neo's heroic efforts as deterministically fated, but freely willed all

the same? Suppose instead that the Oracle did not believe in deterministic fatalism. Perhaps she thought the universe was fundamentally indeterminate and that no facts of the past or present insured any particular way that the future must go. If she believed this, then how did she understand the basis of her own predictions? Maybe in foreseeing Neo's actions, she interpreted them as freely willed and understood her powers to foresee future conduct as completely consistent with the falsity of determinism.[8] The film leaves entirely open which interpretation of the Oracle's beliefs is the correct one.

Consider a very different matter, the status of the enslaved masses. Unlike characters like Neo, Morpheus, and the Oracle, it seems *irrelevant* to ask about what they believe about their own free will and what they might think about fate. They are oblivious to what is taking place outside of the Matrix. Much like the character Truman from the film *The Truman Show*, these poor suckers stuck in those giant wombs are the ultimate illustrations of a very special sort of example used in the free will debate. Incompatibilists are fond of challenging compatibilist notions of control with complicated manipulation cases. The incompatibilists' strategy is to cook up a very troubling scenario in which a person is manipulated into a manner of acting. Of course, what the incompatibilists try to do is make the sort of manipulation so subtle that it is indistinguishable from what ordinary life might be like for you or me. Intuitively the examples are supposed to elicit the reaction that the manipulated person is not free because the source of her action is polluted. It is not she but something else that is the source of her agency. Then the incompatibilists will attempt to argue that a person determined by her past and the laws of nature is no different than a person manipulated in one of these wild scenarios. Hence, the only way that a person like you or me can be free is if she is not

determined. If she is determined, then she is no more free than is a manipulated agent, which is to say that she is not free at all.

These manipulation cases have come to be known as covert non-constraining control (CNC) examples.[9] Compatibilists have two ways in which they can respond to CNC cases. One is to deny that the manipulated agents are unfree. So long as the manipulation is complicated enough, and so long as the manipulation accurately replicates the normal functioning of a person getting through life, then it really is no different than a person being determined. But this is not a problem since the manipulated person is a freely willing one. It is just that the causes of her actions are a lot weirder than the causes of a normally functioning person. Note that this was Cypher's view. In fact, for him the Matrix would afford him more freedom than what was available on that disgusting planet. What did he care what caused his sensation of eating a juicy delicious steak? Real or illusory, he just wanted the damned steak to taste good!

Other compatibilists try to show that there is some significant difference between a causally determined person and a manipulated one. Typically the difference has to do with the history that explains why a person is caused to be as she is. If the causes are of the wrong sort, then she is in some way inauthentic. She is not truly the one engaging the world. Someone or something else is settling for her the values, principles, etc. that she then uses to decide how to act in the world. This, it seems, was the basis for Morpheus's complaint about the Matrix. When he first coaxed Neo, prodding Neo and asking him if he too felt that something about his reality was not right, what Morpheus sought to convey was that human agency within the Matrix was defective; its causal source was designed to settle other goals or needs than

the ones that persons within the Matrix endorsed. Their minds were thus enslaved and so, even if, in a sense, they were "free" within their dream world to do certain things, they were not the source of the goals that their lives ultimately served.

VI.

All of the above reflections indicate the various ways that *The Matrix* openly struggles with the free will debate. But what view of free will is the correct one, and how ought it to be characterized? The philosophical controversy between compatibilists and incompatibilists is one of the perennial problems of philosophy. It will likely remain so. One reason for this is that it is clearly *not* a "no-brainer"! Reasonable minds have differed as to the correct resolution to this problem. And there is no reason to think that this will change any time soon. In fact, one of today's most influential theorists about the controversy has suggested that, at least for certain ways of formulating the problem, the debate between compatibilists and incompatibilists leads to dialectical stalemates.[10] A dialectical stalemate arises when opposing positions within a reasoned debate reach points at which each side's arguments remain reasonable, even compelling, but in which argument runs out; neither can rightly claim decisively to have unseated the legitimacy of the other side's point of view.

I certainly do not know whether the free will problem is ultimately doomed to dialectical stalemate, or whether instead there is some strategy that will be able to settle a reasoned disagreement that is over 2,500 years old. But one point I would like to highlight about this controversy is that it would not have remained a controversial topic, and dialectical stalemates would not have

arisen from it, were it not for the fact that the phenomenology of human experience, as it is for the normally functioning person, does not decisively provide evidence for any one position. It is consistent with how we experience our lives, and how we experience the exercising of our agency, that, in keeping with the incompatibilist position, the control required for free will is illusory, and that we are determined creatures. Or, also in keeping with the incompatibilist position, it is consistent with our experience that the control required for free will is satisfied, and in a way that requires the falsity of determinism. Finally, as the compatibilists allow, it is also consistent with our experience that we do possess free will and that we are determined.

VII.

To its credit, *The Matrix* does not pretend to endorse one point of view about free will. It is neither a compatibilist-friendly, nor an incompatibilist-friendly film. With notable exceptions, the film's reflections on free will mirror the phenomenology of human agency. As is the case in our actual lives, how life is experienced underdetermines the correct answer as to whether the compatibilists or the incompatibilists are correct about free will. I say here "with notable exceptions" since there are clearly aspects of Neo's agency, as well as that of Morpheus's, the other rebels', and the A.I. agents' that most distinctly do not mirror the phenomenology of human agency. It is to these differences that I would now like to turn in closing.

One assumption of the free will debate, shared by all parties to it, is that whatever kind of freedom an agent does possess, whether it requires the falsity of determinism or not, an agent's free will does not consist in her ability to actually cause laws of nature to be false, or to be suspended just in order to

bring about astounding miracles. But within the Matrix, that is, essentially, the sort of control that Neo came to have. Of course, to a lesser extent, so too did Trinity and Morpheus. Indeed, Morpheus even advised Neo to think of the rules of his dream world as mere conventions (rules of a program) that could be bent or just flat out broken. Now some philosophers might want to object here that there is a conceptual problem with describing any rules within the Matrix as both laws of nature and breakable. But this would be splitting hairs at a point at which much more could be gained by reflecting instead upon the power of the thought experiment as it is played out within the film.

Within the history of philosophy, various writers have at one point or another articulated accounts of free will that later were scoffed at and quickly dismissed as fantastical or incoherent or ultimately contradictory.[11] All of these criticisms of these extreme views of freedom might have been on the money, but no philosophical dismissal of the conceptual legitimacy of such a notion of freedom can itself discredit the sort of basis one might have for desiring it. Neo's freedom within the Matrix might seem completely outlandish, merely the stuff of comic books, but the source of its cinematic appeal is that, in a very primitive way, as agents in the world, we all know what it is to bump up against the boundaries of the causally possible. We all understand what a source of liberation it would be if all at once we could act unconstrained by them. Of course, this *is* the stuff that dreams are made of. But to see where our dreams begin often helps us to appreciate both the limits and value of our actual lives.

I shall therefore close with two observations about this extreme sort of fantastical freedom exercised within the Matrix. In section one of this essay I indicated that the freedom of the agents within the Matrix came in degrees,

and that more of it appeared to be more appealing than less. In fact, I suggested that, by the film's end, within the Matrix Neo possessed absolute freedom, and that it rocked. But does absolute freedom rock? We all do value freedom, it appears, and it does look as if it gives most everyone the warm fuzzies. But I propose that *absolute freedom* would not rock, and once had for a while, when exercising it, one would no longer be prepared to exclaim, along with Neo, "Whoa!" This is because the property of rocking found in exercising one's agency comes when one is pressing the boundaries of what she is capable of, pressing the boundaries of the limits placed upon her. Anyone who knows the joy of play understands this. Taking the basketball to the hole, snagging a line drive, pushing one's skis down the steep tight line, nailing a turn on a cycle, or crossing the finish line first with the beat of the pack just behind you, all of this involves the prospect of failure and the demands of an effort of will forced up against the boundaries of what one can do. Absolute freedom would require none of that.

Surprising as it might seem, I propose that a life filled to the brim with absolute freedom would absolutely suck. It would be boring as hell and almost entirely uneventful. Recall the look of utter indifference Neo had on his face when he realized how completely effortlessly he could block Agent Smith's blows in that final face-off. He might as well have been yawning and reading a paper while defending himself: "Ho hum." Imagine if all of one's efforts in life were like this. Contrast this with Neo's intensity and enthusiasm when he still had to work hard to get what he wanted, leaping from a helicopter to save Morpheus, or cart-wheeling through a blaze of bullets and taking out all attackers. How mundane all of this would have been had Neo then been able just to will all of the bullets to stop flying, or Morpheus to stop falling to earth, etc.

Here is a rich irony: Our hankering for absolute freedom, a hankering of a dream world, is something we wish for because we do not have it. Because we bump up against our limits and sometimes fail, we yearn for the power to move beyond those limits. But if we had that power in spades, we'd lose all interest in the activities we find so dear. So it seems that the value of freedom and its place in our lives is partially a function of the manner in which we lack it. It is yet a further credit to a film like *The Matrix* that it instigates such reflections on the value of freedom.

A final speculation will also shed further light on the value we place on freedom. Supposing that Neo could find a way to continue rocking from within the Matrix. Neo faces a fantastic choice. *Should* he work to destroy the Matrix? His absolute freedom is so great within it. Imagine the possibilities. He could be so much in the dream world, have so much, do so much; he could bring such joy to others within it. But knowing what he does about the real world, could he value it, could he take the Matrix seriously? Perhaps you think that Neo should remain within the Matrix where his powers are phenomenal. If instead he attempted to destroy the Matrix, he'd lose all of his powers and have only a dark and barren planet to offer to his liberated human kin. Maybe, like Cypher, they would hate that world and thus resent Neo, seeing him not as a god-like liberator, but as an evil demon dragging them from a relative dream-world utopia into a real-life hell. Even if, for these reasons, you think Neo would do better to remain within the Matrix, acting as a god, trying to do as much good for others as he can, I'll bet that you pause at the thought of it. I myself am unsure what Neo should do, or what I would do if I were he. But if there is something wrong with this option, I suggest that it is at least in part because it would be an inauthentic form of life, *a life that valued a certain kind*

of freedom at the expense of truth, at the expense of real engagement with the actual world. Would this not amount to placing too much value in freedom; would it not amount to valuing freedom at the expense of other worthy elements of life?

When I was a young boy my grandfather, Poppy, took me fishing. I wanted very much that day to catch a trout. I was completely incapable of the task, so Poppy caught one and took it upstream a little way, still hooked on a line. Placing it back in the water, but holding onto the line, he walked it down to me, made as if it was tugging at my pole, and then helped me to "reel it in." I was delighted. So was he. It was only years later that he told me how I came to snag that elusive trout. Suppose that the rest of my life, each fish I caught, I caught only that way, each success of mine was only such a success. Even though Poppy was certainly happy with that little moment of mine, he'd never have wished for me a life of nothing but such shams. To wish merely for an improved life for human kind only *within* the Matrix, even with lots of nifty freedom for everyone within it, I would speculate, if it is wrong, then its wrongness is partially explained by the fact that it is analogous to wishing for all human kind that all of their accomplishments be like Poppy's tying that fish to the end of my pole. It would be nice for a spell, for a moment, in a dream. But we humans want something more. *We* want to catch our *own* fish, and we want to catch *real* fish. When we want something else, we'll go to the movies.

Michael McKenna

Endnotes

1. This claim is meant to be philosophically innocent, simply taking "reality" as the films creators suggested it to be. For proper philosophical scrutiny of the notion of reality as it

pertains to The Matrix, see the essay in this collection by David Chalmers.

2. I shall assume that my reader has seen the film and is familiar with the characters in it, the basic plot, various events that took place, etc.

3. I say that *maybe* Decker is a human being since there is some suggestion in the film that Decker might actually be a replicant and not a human being.

4. For example, in articulating an account of free will, the philosopher Roderick Chisholm wrote:

...if what I have been trying to say is true, then we have a prerogative which some would attribute only to God: each of us, when we act, is a prime mover unmoved. In doing what we do, we cause certain events to happen, and nothing—or no one—causes us to cause those events to happen. (Chisholm, p. 32, cited from Watson, ed., 1982)

Caution should be taken with even this rather extreme view, since Chisholm was not claiming that the sorts of 'miracles' what would allow freely willing and uncaused persons to cause events would amount to miracles that could make walls melt, planes fall from the sky, or bullets to stop in mid air.

5. There is even a controversy amongst compatibilists as to whether or not only the latter notion of control is needed for free will, or whether free will is possible only if both alternative possibilities and actual source conditions are satisfied.

6. For a film that plays with these ideas, see *Minority Report*.

7. This interpretation of the scene fits with Morpheus's subsequent description of how the human race was enslaved. No matter what humans do within the Matrix itself, their conduct is designed to do no more than generate battery juice for the "evolved" artificial intelligences. In fact, it seemed from the film that the level of control that the designers and controllers of the Matrix had over the humans operating within it was not a completely deterministically fated sort of control, but really a sort better suited for no-matter-what-one-does fatalism. This is because people within the Matrix seemed able to do all sorts of different things within certain boundaries. The A.I. creatures cared not a bit. The A.I. intelligences were happy to allow a certain level of social disharmony and chaos amongst the humans within the Matrix. As long as ultimately the outcome was that human lives were lived in the service of creating energy for their artificial intelligence lives, what did it matter to them what the humans did to each other in their dream worlds?

8. The puzzles here over the status of the Oracle's foreknowledge are like those regarding the status of a foreknowing God. If God foreknows all human conduct, does that mean that, by virtue of God's infallible nature, all human conduct is determined? Or is it possible for god to know exactly what any person does or will do even if nothing other than the person herself freely determines what she will do?

9. See Robert Kane, 1996, pp.65-71. . Kane writes:

We are all aware of ...two ways to get others to do our bidding in everyday life. We may force them to do what we want by coercing or constraining them against their wills, which is constraining control or CC control. Or we may manipulate them into doing what we want while making them feel that they have made up their own minds and are acting "of their own free will"—which is covert nonconstraining or CNC control. Cases of CNC control in larger settings are provided by examples like Aldous Huxley's *Brave New World* or B.F. Skinner's *Walden Two*. Frazier, the fictional founder of Skinner's *Walden Two*, gives a clear description of CNC control when he says that in his community persons can do whatever they want or choose, but they

have been conditioned since childhood to want and choose only that they can have or do (p.65).

10. John Martin Fischer, 1994, pp.83-85.

11. A classic example of this is Sartre's notion of radical freedom, which alleged that all persons have freedom with respect to every aspect of reality they confront, every fact of the world. (For an excerpt of Sartre's view, as presented in his *Being and Nothingness*, see the Berofsky collection, 1966, pp. 174-195.)

Suggestions for Further Reading

Books Especially Accessible to an Introductory Audience

Ekstrom, Laura Waddell. 2000. *Free Will*. Boulder, CO: Westview Press.

Honderich, Ted. 1993. *How Free Are You?* Oxford: Oxford University Press.

Wolf, Susan. 1990. *Freedom within Reason*. Oxford: Oxford University Press.

Scholarly Monographs

Berofsky, Bernard. 1987. *Freedom from Necessity*. London: Routledge & Kegan Paul.

Bok, Hilary. 1998. *Freedom and Responsibility*. Princeton, NJ: Princeton University Press.

Clarke, Randy. forthcoming 2003. *Libertarian Accounts of Free Will*. New York: Oxford University Press.

Dennett, Daniel, 1984. *Elbow Room*. Cambridge, MA: MIT Press.

Ekstrom, Laura Waddell. 2000. *Free Will*. Boulder, CO: Westview Press.

Fischer, John Martin. 1994. *The Metaphysics of Free Will*. Oxford: Blackwell.

Fischer, John Martin and Mark Ravizza, 1998. *Responsibility and Control*. Cambridge, UK: Cambridge University Press.

Frankfurt, Harry. 1988. *The Importance of What We Care About*. New York: Cambridge University Press.

Haji, Ishtiyaque, 1998. *Moral Appraisability*. New York: Oxford University Press.

Honderich, Ted. 1988. *A Theory of Determinism*. Oxford: Clarendon Press.

Kane, Robert, 1996. *The Significance of Free Will*. Oxford: Oxford University Press.

Mele, Alfred. 1995. *Autonomous Agency*. New York: Oxford University Press.

O'Connor, Timothy. 2000. *Persons and Causes*. New York: Oxford University Press.

Pereboom, Derk. 2001. *Living Without Free Will*. Cambridge, UK: Cambridge University Press.

Russell, Paul. 1995. *Freedom and Moral Sentiment*. New York: Oxford University Press.

Smalinsky, Saul. 2000. *Free Will and Illusion*. Oxford: Clarendon Press.

Strawson, Galen. 1986. *Freedom and Belief*. Oxford: Oxford University Press.

van Inwagen, Peter. 1983. *An Essay on Free Will*. Oxford: Clarendon Press.

Wallace, R. Jay. 1994. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.

Wolf, Susan. 1990. *Freedom within Reason*. Oxford: Oxford University Press.

Zimmerman, Michael. 1989. *An Essay on Moral Responsibility*. Totowa, NJ: Rowman and Littlefield.

Anthologies

Berofsky, Bernard. ed., 1966. *Free Will and Determinism*. New York: Harper and Row.

Ekstrom, Laura Waddell. ed., 2001. *Agency and Responsibility*. Boulder, CO: Westview Press.

Fischer, John Martin. ed., 1986. *Moral Responsibility*. Ithaca, NY: Cornell University Press.

Fischer, John Martin and Mark Ravizza. eds., 1993. *Perspectives on Moral Responsibility*. Ithaca, NY: Cornell University Press.

Honderich, Ted. ed., *Essays on Freedom of Action*. London; Routledge & Kegan Paul.

Hook, Sidney. ed., 1958. *Determinism and Freedom*. London: Collier.

Kane, Robert. ed., 2002a. *Free Will*. Oxford: Blackwell.

_____. ed., 2002b. *The Oxford Handbook of Free Will*. New York: Oxford University Press.

Lehrer, Keith. ed., 1966. *Freedom and Determinism*. New York: Random House.

O'Connor, Timothy. ed., 1995. *Agents, Causes, and Events*. Oxford: Oxford University Press.

Pereboom, Derk. ed., 1997. *Free Will*. Indianapolis, IN: Hackett.

Schoeman, Freiderich, ed., 1987. *Responsibility, Character, and the Emotions*. Cambridge: Cambridge University Press.

Watson, Gary. ed., 1982. *Free Will*. Oxford: Oxford University Press.

Widerker, David and Michael McKenna. eds., 2002. *Alternative Possibilities and Moral Responsibility*. Aldershot, UK: Ashgate Press.

Especially Influential Articles

A.J. Ayer. "Freedom and Necessity." In Pereboom (1997); and Watson (1982).

Chisholm, Roderick. "Human Freedom and the Self." In Pereboom (1997); and Watson (1982).

Dennett, Daniel. "Mechanism and Responsibility." In Watson (1982).

_____. "I Could Not Have Done Otherwise—So What?" In Kane (2002a).

Edwards, Paul. "Hard and Soft Determinism." In Hook (1958); and Kane (2002a)

Fischer, John Martin. “Responsibility and Control.” In Fischer (1986).

_____. “Responsiveness and Moral Responsibility.” In Pereboom (1997); and Schoemann (1987).

Frankfurt, Harry. “Alternate Possibilities and Moral Responsibility.” In Fischer (1986); Pereboom (1997); and Widerker and McKenna (2002).

_____. “Freedom of the Will and the Concept of a Person.” In Fischer (1986); Kane (2002a); Pereboom (1997); and Watson (1982).

Pereboom, Derk. “Determinism Al Dente.” In Pereboom (1997).

Strawson, Peter. “Freedom and Resentment.” In Fischer and Ravizza (1993); Pereboom (1997); and Watson (1982).

van Inwagen. “The Incompatibility of Free Will and Determinism.” In Kane (2002a); Pereboom 1997); and Watson (1982).

Watson, Gary, “Free Agency.” In Fischer (1986); and Watson (1982).

_____. “Responsibility and the Limits of Evil.” In Fischer and Ravizza (1993); Kane (2002a); and Schoeman (1987).

Wolf, Susan. “Asymmetrical Freedom.” In Fischer (1986);

_____. “Sanity and the Metaphysics of Responsibility.” In Kane (1993); and Schoemann (1987).

For an extensive bibliography, see Kane (2002b).

Plato's Cave and The Matrix

John Partridge

> "Philosophy involves seeing the absolute oddity of what is familiar and trying to formulate really probing questions about it." –Iris Murdoch[1]
>
> "They say about me that I am the strangest person, always making people confused." –Socrates[2]

Imagine a dark, subterranean prison in which humans are bound by their necks to a single place from infancy. Elaborate steps are taken by unseen forces to supply and manipulate the content of the prisoner's visual experience. This is so effective that the prisoners do not recognize their imprisonment and are satisfied to live their lives in this way. Moreover, the cumulative effects of this imprisonment are so thorough that if freed, the prisoners would be virtually helpless. They could not stand up on their own, their eyes would be overloaded initially with sensory information, and even their minds would refuse to accept what the senses eventually presented them. It is not unreasonable to expect that some prisoners would wish to remain imprisoned even after their minds grasped the horror of their condition. But if a prisoner was dragged out and compelled to understand the relationship between the prison and outside, matters would be different. In time the prisoner would come to have genuine knowledge superior to the succession of representations that made up the whole of experience before. This freed prisoner would understand those representations as imperfect—like pale copies of the full reality now grasped in the mind. Yet if returned to the prison, the freed prisoner would be the object of ridicule, disbelief, and hostility.

I. Introduction

Viewers of *The Matrix* remember the moment in the film when Neo is released from his prison and made to grasp the truth of his life and the world. The account above roughly captures that turning point in the 1999 film, and yet it is drawn from an image crafted almost twenty-four hundred years ago by the Greek philosopher, Plato (427-347 B.C.E.). Today the *Republic* is the most influential work by Plato, and the allegory of the Cave the most famous part of the *Republic*. If you know that Socrates was tried, convicted, and sentenced to death by drinking hemlock, or that Socrates thought that the unexamined life is not worth living, you may also know that Socrates in the *Republic* likened the human condition to the state of prisoners bound in a cave seeing only shadows projected on the wall in front of them. Transcending this state is the aim of genuine education, conceived as a release from imprisonment, a turning or reorientation of one's whole life, an upward journey from darkness into light:

> The release from the bonds, the turning around from shadows to statues and the light of the fire and, then, the way up out of the cave to the sunlight...: [education] has the power to awaken the best part of the soul and lead it upward to the study of the best among the things that are.[3]

The allegory of the Cave gives literary shape to Socrates' most fundamental concern, namely that our souls be in the best condition possible (Plato, *Apology* 30a7-b4). Socrates also believed he was commanded by the god Apollo to practice philosophy; it both animated and cost him his life. Yet it is not obvious how philosophical investigation improves the condition of the soul—still less how the Socratic method in particular does so, consisting as it does in testing the consistency of a person's beliefs through a series of questions Socrates asks.

I believe, and will show here, that the allegory of the Cave is part of Plato's

effort to make philosophical sense of Socrates' philosophical life, to link Socrates' persistent questioning to his unwavering aim at what he called the "care of the soul." On this theme of care of the soul, there is a deep resonance between *The Matrix* and Plato's thought in the *Republic*. Like the allegory of the Cave, *The Matrix* dramatically conveys the view that ordinary appearances do not depict true reality and that gaining the truth changes one's life. Neo's movements toward greater understanding nicely parallel the movements of the prisoner in the cave whose bonds are loosened. The surface similarities between the film and the allegory can run to a long catalog. The first paragraph of this essay reveals some of these connections. But there remains a deeper affinity between the two that I shall draw out here, especially in Part IV, having to do with Socrates' notion of the care of the soul.

To see what I am calling a deeper connection between the film and the allegory of the Cave, I begin in Part II by recounting the context in which the Cave appears and the philosophical positions it figuratively depicts.[4] In Part III I compare and contrast the film and the allegory, focusing attention on the difficulty in sorting out deceptive sensory information. Finally, in Part IV I examine the warnings and concessions Plato places in the dramatic spaces of *Republic*. The allegory of the Cave is a strange image, as one of Socrates' friends says (515a4), while Socrates himself confesses that the Cave is not exact (504b5; cf. 435c9-d2).[5] Rereading the Cave after a recent viewing of the film shows that these are not throwaway remarks. *The Matrix* likewise privileges the work that strangeness and calculated vagueness do; Morpheus, after all, cannot show Neo what he most needs to see, but must get him to see for himself something that is difficult to recognize. In this way, *The Matrix* and Plato's Cave are faithful to a central tenet in Socrates' philosophical

examinations: that proper teaching only occurs when students are prepared to make discoveries for themselves. Furthermore, the discovery that is most crucial is the discovery of oneself. Readiness for self-examination is, after all, what makes "care of the soul" possible.

II. Plato's Cave

If Plato's *Republic* has a single unifying theme, it is to show that the life of the just person is intrinsically preferable to any other life. In order to prove this, Socrates is made to investigate the concept of "justice." After an elaborate effort that spans three of the ten books of the *Republic*, Socrates and his two interlocutors discover what justice is. Justice is shown to be a property of a soul in which its three parts do their proper work and refrain from doing the job of another part. Specifically, reason must rule the other parts of the soul. Only under the rule of reason is the soul's harmonious arrangement secured and preserved. Plato glosses this idea memorably by calling such a soul healthy. Just persons have psychic health; their personality is integrated in the proper way.

At the end of Book Four, there is one main gap in the argument: what is the precise role of reason, the "best part of the soul" mentioned in the passage above? There is little to go on at this stage. We know only that the soul in which reason does its job well is called wise, and wisdom is a special kind of knowledge: knowledge of the good. How are we to arrive at this knowledge? What is it like to possess it? What sort of thing is the good? The allegory of the Cave speaks to these questions.[6]

In order to impress upon us the importance of these questions, Book Seven of the *Republic* begins with a startling image of our ignorance. It is the allegory of

the Cave:

> Imagine human beings living in an underground, cavelike dwelling, with an entrance a long way up, which is both open to the light and as wide as the cave itself. They've been there since childhood, fixed in the same place, with their necks and legs fettered, able to see only in front of them, because their bonds prevent them from turning their heads around. Light is provided by a fire burning far above and behind them. Also behind them, but on higher ground, there is a path stretching between them and the fire. Imagine that along this path a low wall has been built, like the screen in front of puppeteers above which they show their puppets . . . Then also imagine that there are people along the wall, carrying all kinds of artifacts that project above it—statues of people and other animals, made out of stone, wood, and every material. And, as you'd expect, some of the carriers are talking, and some are silent. (514a1-515a3)

Many contemporary readers recoil at the awful politics of the Cave. Who, after all, are the "puppeteers"? Why do they deceive their fellow cave-dwellers? Plato has so little to say about them that readers quickly imagine their own worst fears; a totalitarian government or the mass media struck mid- and late-20th Century readers as an obvious parallel to the prisoners who move freely within the cave. But this gets the aim of the cave wrong, I believe, since it deflects attention away from the prisoners bound to the posts. "They are us," Socrates says, and this is what is truly sinister: an imprisonment that we do not recognize because we are our own prison-keepers. Let us turn to examine these prisoners and their imprisonment, specifically by examining the philosophical stakes of their ignorance. Only then will we see exactly why ignorance is likened to imprisonment and alienation.

In the cave, the prisoners can distinguish the different shadows and sounds (516c8-9, cf. e8-9), apply names to the shadows depicting things (cf. 515b4-5), and even discern the patterns in their presentation (516c9-10). To this extent they have some true beliefs. But insofar as they believe that this two-

dimensional, monochromatic play of images—and the echoes reverberating in the cave—is the whole of reality (515c1-2), they are mistaken. Moreover, the opinions they have do not explain why the shapes they see are as they are. They do not know the source of the shadows, nor do they know that the sounds are not produced by the shadows but rather by the unseen people moving the statues (515b7-9).

The possession of a few, small-scale, true beliefs characterizes the condition of all of us, Plato believes. We can distinguish different things, but we lack a systematic, causal explanation of them. To put it loosely, we have, at best, assorted true beliefs about the *what* of things, but a mistaken hold (if any) on the *why* of things. Socrates' search for the definition of justice here, like his search for definitions in other Platonic dialogues, looks like an effort to get at these explanations, to grasp why things are the way they are and, perhaps further, what underlying relationship they have to one another. His questions are part of a search for the essence of things, or what he calls their "form."[7] For Plato, when we possess knowledge of the form of a thing, we can give a comprehensive account of its essence. Without grasp of the form, we can have at best only true beliefs.

A simple example should show what difference it makes to have knowledge of forms.[8] Suppose someone in the cave carries a chair in front of the fire. The bound prisoners see the chair's shadow on the cave wall, and some of them remark, "There is a chair." They are partially correct. If they broke their bonds, they could turn to see the actual chair. In this case their cognitive grip on the chair would be more complete. They would be able to recognize that the shadow was less real than the chair and that the chair is the cause of the shadow.

Ultimately, the physically-real chair is explained in terms of its representation of the form of chair. After all, to have genuine knowledge of a thing it is necessary for our intellects to grasp its form. One might think of the difference this way. A shadow is better grasped when the object casting it is seen. Plato would wish us to see that, in a sense, ordinary objects are like mere shadows of forms. Thus, to grasp objects as fully as possible, one must attain a grasp of its form.

There is a curious complication on the horizon that I shall point out here. It turns out that knowing the form of a thing is not sufficient for gaining a final understanding of that thing. Even to know fully the form of chair, Plato holds, one must know the form of the good.

This does not make sense at first. Recall, the form of the good is what reason ought, ideally, to know, for in knowing it you become wise. Furthermore, knowing the form of the good contributes to your being a just person, since one part of you, reason, is doing its job (and this is what it means for you to be just). Now Plato suggests that grasping the form of the good or the good-itself (the terms are interchangeable; see note 7) is necessary for attaining the best intellectual grasp of *anything* that our intellects can know. The distinctive importance of the form of the good is indicated by two images that immediately precede the Cave: the Sun and the Line, and I will consider them now.

The Sun analogy (507a ff.) reveals the special epistemological role played by the good-itself. Just as the natural world depends upon the sun (for warmth and light), so too the intelligible world depends on the good-itself (508b13-c2).[9] This is the force of the light metaphor. The sun, as Plato puts it, gives the power to see to seers, while the form of the good gives the power to know to

knowers (508e1-3).

In our example of the chair, it is only in virtue of the light produced by the fire above and behind the prisoners that the chair and its shadow are visible. The fire, then, is a condition for our acquiring a more complete true belief about the shadow. But the fire is nothing more than a "source of light that is itself a shadow in relation to the sun" (532c2-3). Out of the cave the sun represents the good-itself. The good-itself illuminates the true, intelligible world of ultimate reality, and in this way, the form of chair relies on the form of the good for its intelligibility. The good-itself is the most preeminent item in the universe. It is both an object of knowledge and the condition of fully knowing other objects of knowledge.

Plato is not finished with his specification of the role played by the form of the good. He goes on to suggest that the good-itself nourishes the being of intelligible things in a way analogous to the sun nourishing organic life. For this unusual idea we have some help from the Line image (509d ff), the most obscure of the three images. Imagine a vertical line dividing two realms—physical reality and intelligible reality—into unequal spaces. Each realm is then subdivided in the same uneven proportion as that which separates the physical and intelligible world. To take only the smaller, bottom portion of the line, we find the physical realm divided between actual, physically-existing items and their ephemeral copies (e.g., reflections in water, shadows, and artistic depictions). In the Cave, this is the distinction made between the chair and its shadow. And so too the Line presses us to think that the physically real objects perceived by our senses are, in effect, shadows—pale, diminished or distorted copies of something more real.

The Line offers a ranked order of Plato's ontology according to which the degrees of reality and being of a particular class of things increases as you go up the line. The higher up the scale, the more real the items become; and since the form of the good is the most real item in all of reality, it is located at the very top of the Line, just above the forms. Things lower on the line are derivative and owe whatever reality or being that they have to the things above them. Physical objects are, metaphorically, nourished by their corresponding forms. They depend for their very reality, not just their knowability, on the perfect, eternal Forms existing in the intelligible realm.

One clear implication of the Line is the metaphor of ascent. The Cave exploits it as well: the upward escape from the cave represents the difficulty of gaining ever more abstract knowledge while not relying on information gathered by the senses. By connecting the three images together we discover that the human condition is abject: we see only the most downgraded forms of reality (image, shadows) and are as far from the sun (the good-itself) as we can be. This is what it is to be ignorant of the truth.

But to see why our alienation from what is genuinely good makes a difference in our lives, there is one more feature of the good-itself that deserves attention. Whatever exactly the form of the good is, it serves as a paradigm or model, and it has a remarkable effect on those who grasp it. As Socrates says of fully-educated philosophers near the end of Book Seven, "once they've seen the good-itself, they must each in turn put the city, its citizens, and themselves in order, using it as their model (*paradeigmati*)" (540a8-b1). This was anticipated in a longer passage in which the philosopher, by means of studying the "things that are" (500b9), acts as a craftsman (cf. 500d6), or a "painter using a divine model (*paradeigmati*)" (500e3-4). Not only do physical things

take on the qualities they have through a process of copying, reflecting or imitating the forms, so too we can take on goodness through intellectual contact with the good-itself.[10] By coming to understand the good-itself, we become like it. In short, we become good.

We can see now why being just depends on knowing the form of the good. Reason's rule affords the soul the opportunity to study and therein to become like the good-itself, that is, properly proportioned, well ordered, healthy. Finally, once this knowledge is acquired, and the self is transformed, one becomes productive.[11] Those who gain knowledge of the good-itself are capable of crafting virtues in their souls and in the souls of others, and they can paint divine constitutions for cities. This is what enables Plato to put words into Socrates' mouth that, were he on Aristophanes' stage, would have returned thunderous laughter:

> Until philosophers rule as kings or those who are now called kings and leading men genuinely and adequately philosophize, that is, until political power and philosophy entirely coincide . . . cities will have no rest from evils, Glaucon, nor, I think, will the human race. (473c11-d6)

III. Plato's Cave and *The Matrix*

There are no forms in *The Matrix*, and thus our epistemic and metaphysical circumstances in Plato's *Republic* look very different from those in the film. The world inside the cave is a diminished one, a shadow or reflection of the real, but broadly continuous with the true world. Even though there is a marked difference between the sensible and intelligible realms viz. method, epistemic certainty, and metaphysical reality, on Plato's view the sensible is somehow derived from the intelligible. Thus, for Plato, our speaking and thinking in the cave is not meaningless, and some of our opinions are true, in spite of our ignorance of the deeper causes of things.

In *The Matrix*, by contrast, the two worlds are far less continuous with one another. The real world is profoundly dystopian, and the substance of lives inside the Matrix is supplied in mental states almost entirely cut off from this reality. (Ironically, the real world in *The Matrix* is very like the world inside the cave.) In spite of its realism, the world inside the Matrix is not a copy of the real world but is a simulation. Nevertheless, there is at least one continuity between the real world and the computer-simulated world: your body. Owing to an unexplained principle, called "residual body memory," your body looks the same to you and to others in both worlds. And you are able to retain your memories of one world when you are in the other and when you return back to the first. (This means that Cypher will have to have his memories of the time spent outside the Matrix removed if he is to return to the illusion of reality inside the Matrix.)

Since the real world and the simulated world are worlds in which the senses receive information, the practical problem is not that they are discontinuous, but that they are indiscernible. This is part of the initial difficulty for Neo since he cannot determine which sensory information is genuine and which false. Although he (and the viewer) settles this question soon enough, a skeptical worry remains in the wake: how can he ever be sure his sensory information is truthful if there is no certificate of authenticity on his experiences?

Suppose Agent Smith creates a program that launches right when Neo picks up a phone within the Matrix. Instead of being whisked back aboard the ship, Neo's consciousness is supplied with a computer-generated experience of the interior of the Nebuchadnezzar, and of course he believes he has successfully exited the Matrix. Such a trick might enable Agent Smith to obtain compromising information about the Nebuchadnezzar and its crew or, worse,

the passwords for Zion.

It is hard to imagine how Neo might see past Agent Smith's ruse, especially if he only had a few moments to figure things out. Would Plato's freed prisoner fair better? Recall, Plato urges us to regard the sensible world as unreliable, no matter the source of our information about it.[12] We must adopt a different method for apprehending the truth of things. This is, of course, not nearly as simple as it sounds, nor is it obviously helpful; after all, what we are to grasp is the intelligible world from which our ordinary, sensible world is copied, not the sensible world itself. The reward is that once you grasp the forms in the intelligible world, you would be an expert in discriminating items in the sensible world (cf. 520c1-6). This doesn't mean you'd never be mistaken, however; rather, you would simply be the best sensible world discriminator there could be. Therefore, in the case where Agent Smith launches his deceptive program, the only advantage the freed prisoner might have is slight: a general unease about all sensory information. Since the ordinary world is too murky and ever-changing to permit genuine knowledge of it, our awareness of this mutability should assist us in determining which of our beliefs were relatively more reliable.

It seems that the metaphysical differences between Plato and *The Matrix* do not prevent them from telling a roughly similar story about the epistemological unreliability of the senses and the need to abstract from the senses in order to gain genuine knowledge. In fact, we find Neo at the end of the film doing more than simply bending the laws of physics with the Matrix. He has, it seems, stepped almost entirely out of that very world itself. He does not, however, appear in two places at once, but his destruction of one of the Agents, and his ability to fly, suggest that the laws of physics are more than merely bent.

Where Plato's dialogue and *The Matrix* agree most is in drawing out the enormous psychological difficulty in calling the world into question and the ethical dimensions of failing to do so. Neo and Plato's freed prisoner must accept truths about themselves (namely, that their lives have been unreal) before they can acquire deeper knowledge about fundamental truths. To achieve this, both Neo and the freed prisoner need the shocking demonstration that the senses are inadequate and that they can be systematically deceived. Both then undertake an introspective turn to discover the truth, and must take steps to disregard knowledge derived from the senses.

This is the point to ask, finally, what knowledge Neo attains that operates in him like the knowledge of the Platonic form of the good. What does Neo know only after great difficulty but whose truth is fundamental? What object is grasped by Neo's intellect that he understands to be the condition of his knowing anything else? What knowledge enables him to be productive, to be a savior of himself and others? It is nothing more than proper self-understanding. In both *The Matrix* and in the Cave, there is a single item the knowledge of which makes the knower more integrated and more powerful, and for Neo it is self-knowledge.

Ought we to see Neo as adhering to the letter of Socratic self-examination and care of the soul? Only at high-altitude will a perfect connection be visible. For Neo's enlightenment is ultimately about his own specific path and role. Socratic care of the soul involves self-knowledge, but the parts of yourself that are peculiar to you, that make up your individuality, are not relevant.[13] Since the prisoners in the cave have only dim self-awareness (they see only the shadows of themselves [515a5-8]), it might seem that release involves getting the right

beliefs about oneself. But the very abstractness of the knowledge that Plato prizes, which is very unlike the specificity of the knowledge that Neo eventually gets (namely, that he is the One), suggests that the self-knowledge the prisoners need is neither the end of their search nor even the proper beginning.

In other dialogues Socrates was made to endorse the idea that knowledge was in you, that a kind of introspection aided by proper questioning could elicit true beliefs. But these are not truths that are *about* you, rather they are truths that are *in* you. Neo's case is different. The truths he must grasp are both in him and about him. The film reveals furthermore how he must demonstrate and experience his capabilities before he is able to believe entirely that he possesses them. And when he believes in himself at last, his capabilities are further enhanced. *This* result is produced neither by the method nor the aim of Socratic care of the soul.

Most fundamentally, the film and the allegory share a pedagogical conceit. Both hold that in teaching the most basic truths, there is an important role for a strategic strangeness and the confusion it produces. The allegory of the Cave puzzles Socrates' audience, yet as it hooks them, the Cave provides only the outline for solving the puzzle. Might Morpheus be doing the same? Might Morpheus, like the allegory, act as a kind of Socratic teacher, urging Neo toward self-understanding and care for his soul?

IV. Socratic Education in the Cave and *The Matrix*

To see to what extent this is so, I want now to return to a remark by Socrates' friend, Glaucon, that the cave and its prisoners are "strange" (*atopon . . . atopous*, [515a4]). The remark is important because it indicates that the image

is operating on its audience in a particular way, one that Plato elsewhere gives us reason to believe is significant. Prompting someone to recognize strangeness, something being out of place (*atopia*), is how the Socratic method achieves one of its aims. This can occur when Socrates asks one of his deceptively simple questions. But it can also occur when he professes ignorance, or when he is silent. Similarly, Plato's allegory of the Cave describes what our ignorance is like in stark images and what it would be like to become educated; it says nothing about what starts the process of becoming educated.[14] Of course, the imprisonment is metaphorical, as is the release. Pressing for specific details is to demand too much of the image. By refusing to say precisely how *this* prisoner is freed, Plato retains the openness of his allegory.[15]

What are we to say about *The Matrix*? On the surface, it appears the *The Matrix* departs from the allegory. First of all, it gives answers to the question above, for it is Morpheus who frees Neo, and Morpheus chooses to free him because there is something particular about Neo that recommends his release. Yet, on closer inspection, Neo's early encounters with Morpheus produce the same kind of confusion that Socrates produces in his interlocutors. Neo receives strange communications via computer ("wake up, Neo,"[16]) to follow the white rabbit he soon sees on a tattooed shoulder. These odd messages disrupt Neo's expectations of the world, especially his need for control over his life and his facility with computers. Another disruption comes when Neo swallows the red pill. This drug quickly begins to alter his perception of the stability of the world inside the Matrix.[17] Taken together, the computer messages uncannily anticipate what is about to happen, while the pill calls into question his grasp of what is now happening. This surely prepares Neo to

accept the truth that everything that has already happened is an illusion.

If we suppose that Morpheus asks the right questions, and supplies the right drugs, it is still the case that Neo has to recognize the questions and accept the drugs. Neo proves to be a particularly apt pupil. Indeed, there are features of Neo's life that might explain how he begins to see the falsity of the world inside the Matrix. Neo is an accomplished hacker who would have the best chance of anyone to discover that the whole of his experience is itself nothing more than highly-sophisticated computer code. He is also living a double life. He works as a software engineer perhaps to maintain a steady income, perhaps as cover for his underground activities. Maybe playing the role of an office worker affords him a sense of the absurd that makes it easier to believe that his life is hollow. Insomnia might work for this purpose as well. Besides, who hasn't had the gut feeling Neo has that "there is something wrong with the world"?

Of course, one of the themes of the film is Neo's struggle to accept his role as the One, the savior of humanity. He is the subject of a number of prophecies made by the Oracle.[18] In fact, he is the only person whose prophecy does not refer to someone other than himself. He only accepts his true nature well after the series of strange clues Morpheus presents to him and the confusion this produces in him. Ultimately, he must experience first-hand his fitness for the special role that the others urge him to perform.

In this way, Morpheus can be seen as a Socratic gadfly, stinging Neo to take the first steps he needs in order to discover the truth on his own. Similarly, Plato's sketch of the role played by the form of the good only points the way to the complete answer that Plato would have us seek out. In this way, Plato draws the reader to think for him or herself in the same way that Socrates wished his interlocutors to feel the sting of the realization of their ignorance as

a motivation to join him in inquiry and care of the soul.

The allegory of the Cave issues a pointed challenge: in what way are we living lives of diminished prospect, resting content with our knowledge, failing even to ask the right questions? These are precisely the questions Morpheus puts to Neo. And like Morpheus, Plato's pessimism about the human condition gives way to an optimistic view of the power of education to liberate anyone:

> Education isn't what some people declare it to be, namely, putting knowledge into souls that lack it, like putting sight into blind eyes . . . Education takes for granted that sight is there but that it isn't turned the right way or where it ought to look, and it tries to redirect it appropriately. (518b7-c2, d5-7)

John Partridge

Endnotes

1. "Literature and Philosophy: A Conversation with Bryan Magee" in *Existentialists and Mystics: Writings on Philosophy and Literature* (New York: Penguin, 1998), 8. Originally published in Magee, *Men of Ideas* (Oxford: Oxford University Press, 1978).

2. Plato, Theaetetus 149a8-9.

3. Plato, *Republic* 532b6-8, c3-6. What I have dubbed "education" in the brackets is specifically the study of mathematics, geometry, astronomy, and harmony. When properly pursued, each discipline involves abstraction from the senses, and is "really fitted in every way to draw one towards being" (523a2-3). These disciplines prepare our minds for the most important discipline, dialectic, Plato's term for the right kind of philosophical examination.

Hereafter I include citations to the "Stephanus pages" of the *Republic* in the text. Stephanus pages may be found along the vertical margins of most translations of the Republic. For example, "527d6-e3" refers to a passage beginning on Stephanus page 527, section d, on line 6 of the Oxford Classical Text. The translation I cite here is by Grube/Reeve (1992), which is also found in Cooper (1997).

4. I shall refer to the philosophical positions advocated by the character Socrates as Plato's, though this scholarly convention is under attack in some quarters. Plato never appears in the Republic or any other dialogue (save for the Apology, and he does not speak there). Thus some scholars find it presumptuous to fob off the character Socrates' views onto Plato; would we automatically assume that Ian Fleming took his martini shaken, not stirred, in the manner of his fictional agent? Of course, more is at stake in the first case than getting a drink order wrong, but this is true largely because other assumptions normally accompany the identification of Socrates' utterances with Plato's considered philosophical views. One worry is

that this identification narrows the range of answers we might give to the question why Plato wrote dialogues. Another worry is that it may distort our understanding of what Plato took an adequate philosophical theory to be.

5. Contemporary readers generally agree with Socrates. Some refer to "the treacherous analogies and parables" (Cooper [1977], 143) as "over-ambitious" and "overloaded" (Annas [1981], 265; 252, 256). Much ink has been spilled in the effort to provide a consistent, plausible philosophical interpretation of the images in the *Republic*.

6. I say "speaks to" because the Cave is only part of a generally sketchy account of the nature of the good. Socrates disclaims precision, warning us that his talk about the good is schematic (504d6-8) and fuzzy (cf. 504d8-e3); a shortcut to the truth of things (cf. 504b1-4; 435d2). Given his lack of knowledge about the good (505a4-6, 506c2-3, d6-8), the most Socrates can do is provide stories, not reasoned accounts. This, at least, is the stated rationale for why he gives "the child and offspring of the good" (507a3-4) rather than a fully articulated, rationally defensible account.

Socrates' disavowal of knowledge does not mean that he is completely ignorant. Most obviously, he knows enough to know that he does not know. He also knows that knowledge of the good is important to have (505a6-b4), and what method must be used to get it: dialectic (532a1-d1). Moreover, he provides a formal account of the good, saying it is the chief or ultimate end to all our actions (cf.: "Every soul pursues the good and does whatever it does for its sake" 505d11-e1). And with this premise, he rules out rival attempts to spell out the formal account, arguing against pleasure and knowledge as candidates for a substantive account of goodness itself (505b5-d1). Finally, he seems capable of saying more than he says here, though we cannot be sure that he takes himself to be able to give something more secure than images and other "offspring" (cf. 506e1-3).

7. See 507b5-7. The essence of good things is called, variously, the good-itself (506d8-e1, 507b5) or form of the good (505a2, 508e2-3). This item is really what reason is attempting to grasp; not what is good for me, nor what is 'a good x', but something that is good in and of itself.

8. It is notoriously difficult to count the population of forms, and we cannot be certain that Plato thought there was a form of chair. Reeve's comment (on whether there is a form in the intelligible world for every group of things in the sensible world to which a single name applies) is useful for the general question of how many or what sort of forms there are. "Assumptions are one thing; truths are another. Thus forms are assumed with ontological abandon, but the only ones there really are are those needed by dialectical-thought for its explanatory and reconstructive purposes. Ordinary language is the first word here, but it is not by any means the last word" (1988, 294). Will there be a last word? According to one commentator writing at the beginning of the last century, even what Plato meant by the forms "is a question which has been, and in my opinion will always be, much debated" (Adam [1902], 169).

9. The intelligible world is Plato's way of referring to the class of things that can be known by the mind alone and that are imperceptible to the senses. A list would include mathematical or logical truths and geometrical items, as well as the vaunted forms. (The types of study that yield knowledge of items in or aspects of the intelligible world are mentioned in note 3 above.)

10. "Instead, as he looks at and studies things that are organized and always the same, that neither do injustice to one another nor suffer it, being all in rational order, he imitates them and tries to become as like them as he can. Or do you think that someone can consort with things he admires without imitating them? . . . Then the philosopher, by consorting with what is ordered and divine . . . himself becomes as divine and ordered as a human being can" (500c2-7, c9-d2).

On some ears, this kind of talk encourages mysticism, or the view that the good-itself has occult qualities. But we do well to remind ourselves that dialectic is the only route to grasping the good-itself, and that dialectic is studied only after ten years of mathematics, geometry, astronomy, and the like (537b-c). Indeed, Cooper has argued that we think of the good-itself "somehow or other as a perfect example of rational order, conceived in explicitly mathematical terms" ([1977], 144; see also Kraut [1992]). Again, it is intellectual grasp—not oneness with or absorption into the good—that we are striving to attain.

11. Plato's *Symposium* famously stresses the fertility of the philosopher who has grasped the forms (212a-b).

12. For this reason, Plato might appreciate the irony of Morpheus stressing, again and again, that Neo must see for himself in order to understand. Plato would regard Neo's transformed conception of reality partial at best since Neo is not called upon to regard all sense impressions as false or diminished, only those that have the wrong source.

13. Annas (1981), 257-59, makes this point when she compares Plato's allegory to Bertolucci's 1970 film, *The Conformist*.

14. In the allegory, the prisoner's chains are removed but Socrates is silent on who or what removes them. Here are his words: "Consider, then, what being released from their bonds and cured of their ignorance would naturally be like. When one of them was freed and suddenly compelled to stand up, turn his head, walk, and look up toward the light, he'd be pain and dazzled and unable to see the things whose shadows he'd seen before" (515c4-d1). The Cave depicts an astonishingly thorough imprisonment. Throughout, Plato remarks on the difficulties that the freed prisoner meets with on the way out of the cave. Given this detail, it is not unreasonable to expect an account of precisely what sort of prisoner it is who begins to question whether the cave contains the whole of reality, or precisely what circumstance prompts his inquiry. Does the prisoner find the play of shadows internally inconsistent? Or does one or more of the unbound prisoners decide to remove the bonds? We are not told.

15. Moreover, the freed prisoner is referred to generically by the indefinite pronoun "someone" (*tis*); if we wish for specifics, we miss the generality that Plato intends, for his point surely is that *anyone* could escape the bonds of ignorance.

16. The film surely intends us to read the figurative sense of this expression alongside the literal one, and it may be Morpheus' hope that Neo reflects on the figurative meaning as well. After all, one of the other messages that appears on his screen—"knock, knock, Neo"—is consciously riddling. It invites the question, "who's there?"

17. Although the aim of the pill is to assist in locating Neo's body, the suggestion of a psychoactive effect on him is unmistakable.

18. The Oracle eventually tells Neo "what he needed to hear," namely that he is not the One. This inverts the account of Socrates' oracle as Plato portrays it in the *Apology*. First, Socrates does not hear the oracle directly but relies on Chaerephon's report that "no one is wiser than Socrates." Second, Neo's reluctance to believe that he is not in control of his actions requires that the Oracle tell him something false. This Neo is happy to hear, and thus he has no motive for questioning it; it is eminently believable that he is not their long-awaited savior. By contrast, Socrates' oracle tells him something true but whose unlikely implications must be carefully interpreted through testing and questioning.

Works Cited / Suggestions for Further Reading

Adam, James. *The Republic of Plato*. 1902. 2nd Ed. Cambridge: Cambridge University Press, 1963.

Annas, Julia. *An Introduction to Plato's Republic*. Oxford: Clarendon Press 1981.

Cooper, John M. "The Psychology of Justice in Plato." *American Philosophical Quarterly* 14 (1977): 151-57.

Cooper, John M, ed. *Plato: Complete Works*. Indianapolis, Indiana: Hackett Publishing Co., 1997.

Grube, G.M.E, trans. *Plato: Republic*. 2nd Ed. Rev. C.D.C. Reeve. Indianapolis, Indiana: Hackett Publishing Co., 1992.

Kraut, Richard. "The Defense of Justice in Plato's Republic." In *The Cambridge Companion to Plato*, edited by Richard Kraut. Cambridge: Cambridge University Press, 1992, 311-37.

Kraut, Richard, ed. *Plato's Republic: Critical Essays*. Lanham, Maryland: Rowman & Littlefield, 1997.

Reeve, C.D.C. *Philosopher-Kings: The Argument of Plato's Republic*. Princeton: Princeton University Press, 1988.

CONTRIBUTORS

David Chalmers is Professor of Philosophy and Director of the Center for Consciousness Studies at the University of Arizona. He is author of *The Conscious Mind: In Search of a Fundamental Theory* (Oxford University Press, 1996). He is especially interested in consciousness, artificial intelligence, metaphysics, and meaning.

Julia Driver currently teaches at Dartmouth College. Her main research interests are in ethical theory and moral psychology, and she has published a book (*Uneasy Virtue*, Cambridge) and a variety of articles in the area of normative ethical theory. She is co-editor of the Normative Ethics section of *The Stanford Encyclopedia of Philosophy*. (http://plato.stanford.edu)

Hubert Dreyfus was educated at Harvard and teaches philosophy at the University of California, Berkeley. His research interests bridge the Analytic and Continental traditions in 20th-century philosophy. He has written books on Heidegger (*Being-in-the-World*, MIT Press), and on Artificial Intelligence. (*What Computers (Still) Can't Do*, MIT Press). Dreyfus recently published *On the Internet* (Routledge), and is working on a book with Charles Taylor tentatively entitled, *Retrieving Realism*.

Stephen Dreyfus is a graduate of Video Symphony, just beginning his professional career as a digital film editor. He has worked as an Assistant Editor on several independent films. A long time amateur philosopher, he is always looking for new and interesting ways to bring his surrealist stories and ideas to the entertainment world. His email address is lorde_red@lycos.com.

Frances Flannery-Dailey received her Ph.D. from the University of Iowa, and is Assistant Professor of Religion at Hendrix College in Conway, Arkansas. She teaches courses in Bible, Religion and Culture and Judaism. Her main area of research is apocalypticism in early Judaism (300 B.C.E.-200 C.E.), and she is currently writing a book for Brill Publishers entitled *Dreamers, Mystics and Heavenly Priests: Dreams in Second Temple Judaism*.

Christopher Grau was educated at Johns Hopkins University and New York University. In addition to editing the "Philosophy & *The Matrix*" section of *The Matrix* website, Chris is Assistant Professor of Philosophy at Florida International University in Miami. He has previously taught at Dartmouth College, Johns Hopkins University, Brooklyn College, and the University of Maryland, Baltimore County. His current research involves the ethical ramifications of theories of personal identity.

Richard Hanley was educated at Sydney University and the University of Maryland, College Park. He is the author of *The Metaphysics of Star Trek* (reprinted in paperback as *Is Data Human?*), and is co-editor of the forthcoming *Blackwell Guide to the Philosophy of Language*. He works on metaphysics, philosophy of language, and ethics, and dabbles in time travel fiction; and is gainfully employed in the Philosophy department at the University of Delaware.

Colin McGinn was educated at Oxford University. He has written widely on philosophy and philosophers in such publications as the *New York Review of Books*, the *London Review of Books*, the *New Republic*, and the *New York Times Book Review*. McGinn has written fourteen books, including *The Making*

of a Philosopher; The Mysterious Flame; The Character of Mind; Ethics, Evil and Fiction; and the novel *The Space Trap*. He is currently a Professor of Philosophy at Rutgers University.

Michael McKenna received his Ph.D. from the University of Virginia in 1993. He is an Associate Professor of Philosophy in the Department of Philosophy and Religion at Ithaca College. McKenna has published various articles on the topics of free will and moral responsibility. He is currently working on a book devoted to a communication-based account of morally responsible agency. McKenna teaches courses in metaphysics, moral and political philosophy, and philosophy in film.

John Partridge is Assistant Professor of Philosophy at Wheaton College in Norton, Massachusetts where he teaches courses in aesthetics, philosophy and literature, and the philosophy of the emotions. He received a Ph.D. from Johns Hopkins University and has published articles on Plato in *Ancient Philosophy* and *Skepsis*.

James Pryor was educated and teaches philosophy at Princeton. He has published on the epistemology of perception, and works primarily on philosophical issues concerning the mind and knowledge.

Iakovos Vasiliou is Associate Professor of Philosophy at Brooklyn College, City University of New York. He has previously taught at Cornell, Johns Hopkins, and Georgia State University. He has published articles on Plato, Aristotle, and Wittgenstein and is currently working on a book on moral epistemology in Plato and Aristotle.

Rachel Wagner was educated at Wake Forest University and the University of Iowa. She is currently a Visiting Assistant Professor of Religion at Southwestern University in Georgetown, Texas. At Southwestern, she teaches comparative courses in Islam, the Problem of Evil, Religion and Film, and Religion and Literature, incorporating materials from many different religious traditions. Her primary training is in Western Religions and biblical studies, and her dissertation is based on William Blake's epic poem *Jerusalem*.

Kevin Warwick is a Professor of Cybernetics at the University of Reading, UK. His book *In the Mind of the Machine* gives a warning of a future in which machines are more intelligent than humans. In 1998 and 2002 he received surgical implants which shocked the scientific community. The second of these linked his nervous system to the internet. His experiments are reported on in his autobiography *I, Cyborg*.