What's So Bad about Living in The Matrix?

David Pincus

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