Member-only story

No, We're Not Living in a Simulation

Richard Johns
Published in Predict
19 min read · May 26, 2020

A Sim could not understand the concept of a being a Sim

Image: Theo Johns

> *… there being nothing so inconceivable as to say that matter comprehends itself. It is not possible for us to comprehend how it would comprehend itself.*
>
> *~ Blaise Pascal*

Can a Sim understand things?

In recent years many people have raised the idea that the supposed reality we are experiencing might in fact be a simulation, running on a computer that exists in the true (or "base") reality. There are many arguments for this, some more serious than others, but all such arguments make an assumption that I reject, namely that it is at least *possible* to create conscious, rational beings within a computer program. Note that the simulation hypothesis is quite different from the plot of the *Matrix* in this respect, since most of the people in the Matrix are flesh-and-blood human organisms, existing in the base reality, who merely get their sense experiences from the program. The simulation hypothesis, on the other hand, says that we humans are just "Sims", i.e. entities that are part of the simulation, created by the computer code, like non-player characters in a video game.

A lot of people have trouble believing that it's possible to create a conscious and rational person simply by running a program, no matter how complex the program is. I think they're right to be skeptical. In fact, I believe the very idea of a simulation being an actual rational person is implicitly contradictory — as I'll explain in this article. Before we get to that, however, we might wonder why anyone would think that it *is* possible to create a person within a computer program. What's the reason for that?

The short answer is that, over the past couple of centuries, scientists and philosophers have come to the firm opinion that human beings are complex machines, made only of physical matter. (The main alternative idea, which added an immaterial soul to the physical body, was judged to create more problems than it solves.) Also, after computers were invented, it was found that any machine can be simulated (with any degree of precision you want) in a computer program, as a so-called "virtual machine". The virtual machine is equivalent to the physical machine in its 'function', which roughly means that it responds the same way, given the same stimulus. For example, the TRS-80 computer from 1977 can be 'emulated' by a program running on a modern computer, which then behaves exactly like a TRS-80, complete with blocky graphics and monochrome display. Now, supposing that humans are physical machines, they too can be adequately simulated by a computer program. Whatever the person does in their physical environment, the Sim will do if its virtual environment prods it in the same way. But what *mental life* would such a virtual human have? Would they also be

conscious and rational? Most philosophers and AI researchers say yes, there would be no cognitive difference between the physical machine and its virtual counterpart. They're functionally equivalent, and hence mentally equivalent as well — this is a view of the mind known as 'functionalism'. The main reason to accept functionalism is that it has fewer problems than any alternative materialist view.

Critics of functionalism have argued in various ways that virtual machines can't be real people — and many of these arguments would also show (if they work at all) that physical machines can't be real people either. When these critics say that a person cannot exist in a simulation, they usually have one of two reasons for this. One reason is that people are *conscious*, and the other is that people have *understanding* (or, as philosophers say, 'intentionality'). The argument from consciousness is basically that machines (both physical and virtual) can be fully understood, at least in principle. Understanding a machine, however, involves concepts of structure and dynamics that seem not to be logically related to many of the concepts we get from our conscious experience. From sense experience we have 'phenomenal' concepts like red, sweet-tasting, smelling of rotten fish, and so on, which seem unconnected to physical concepts like spatial position, velocity, wavelength, electric charge, and so on. If humans were purely physical, then a complete physical description of a person would have to convey all true information about them, including their conscious experiences. On the contrary, however, any physical description (no matter how complex and detailed) seems to say nothing about *what it feels like* to be a human being.

This argument from conscious experience, or 'phenomenal qualities', has been discussed at great length over many decades, and I'll say no more about it. Here I want to talk about a new argument, first published earlier this year and virtually unknown at present. This new argument finds a logical contradiction in the claim that a machine (either physical or virtual) can have rational understanding to the same degree as a human.

When we talk of humans as rational beings, there are two main components to this idea of rationality. The first is that humans are able to *understand* the world, and form *beliefs* about the world (as well as speculate, have desires and so on). Believing and understanding involve forming internal states, within one's brain, that are *about* the world, in that they represent things in the world as being a certain way. For example, if

one believes that (or imagines or speculates that) Mars has 3 moons, then somewhere in one's mind there is a representation of Mars having 3 moons. This is an example of what philosophers call an *intentional* mental state, one that is *about* some state of affairs in the real world, or some possible state of affairs. Having a headache, by contrast, is not an intentional state, as it doesn't involve any proposition or claim about the world. Feeling thirsty isn't an intentional state, but wishing I had a bottle of cold beer in my hand is. The second component to being rational is that one's belief system obeys logical norms, such as logical consistency, deductive closure and the axioms of probability, but this plays no role in my argument.

Humans have intentional states, and each intentional state includes a so-called 'proposition' as its 'intentional content'. For example, wishing I had a cold beer in my hand has the (false) proposition *I have a cold beer in my hand* as its intentional content. If we assume that every human is a virtual machine, running on a finite computer, the state of a human mind at a given moment must be defined by some finite set of parameters, each with a finite set of possible values. Also, knowing the values of these parameters, together with an understanding of the program itself, would provide a sufficiently clever programmer with a full understanding of what proposition the human is thinking when in that state. I call this the *mind-reading thesis*, that a godlike programmer, with unlimited mental ability, could read the minds of all the virtual humans within the earth simulation, in something like the way that some characters in the *Matrix* movie can read the green Matrix code raining down the screen, and infer what is happening within the simulation.

The contradiction that arises from the simulation hypothesis uses a diagonal argument, a technique developed by mathematicians over a hundred years ago. The way diagonal arguments work is explained in the next section.

Diagonal arguments

Here's the simplest kind of diagonal argument, based on an argument given by mathematician Georg Cantor in 1891. Suppose that Ralph, a friend of yours, tells you that there are just five different 5-bit sequences. In fact, he has assigned a code number (from 1 to 5) to each of the 5-bit strings he recognizes, as shown in this table:

Code number	Binary sequence				
1	1	0	1	0	1
2	0	1	0	1	0
3	1	1	0	1	1
4	0	0	1	0	0
5	1	1	0	0	1

We know of course that Ralph's table is woefully incomplete: there are in fact 32 such strings, not just 5. A neat way to show that the table is incomplete is to use a *diagonal argument*, which involves constructing a 'diagonal' sequence of bits as shown below:

Code number	Binary sequence				
1	**1**	0	1	0	1
2	0	**1**	0	1	0
3	1	1	**0**	1	1
4	0	0	1	**0**	0
5	1	1	0	0	**1**

In other words, the diagonal sequence for this table is 11001. This sequence is also in the list itself, at the bottom, although this won't be the case for most lists.

A more useful sequence is the *inverted* diagonal, 00110, which is made by changing every bit of the diagonal string. This cunning definition means that it *cannot* be on the list, no matter which five strings Ralph includes, as it differs from every sequence on the list in at least one bit. Since the inverted diagonal is missing from the list, we have proved that the list is incomplete.

Ralph might also try to list all the sets that can be formed from the numbers 1 through 5, such as the sets {1, 3, 5} and {2, 3}. This problem is actually equivalent to the previous one, since one can interpret each 5-bit string as specifying the membership of such a set, in the obvious way. Converting each binary sequence in Ralph's table above to a set, we obtain this table.

Code number	Set	Inclusive?
1	{1, 3, 5}	Yes
2	{2, 4}	Yes
3	{1, 2, 4, 5}	No
4	{3}	No
5	{1, 2, 5}	Yes

The third column indicates which code numbers are 'inclusive', in the sense that they are members of their corresponding set. The set of all the inclusive code numbers, i.e. the set {1, 2, 5}, is equivalent to the diagonal sequence 11001 above. The inverted diagonal set is the set of all the *non*-inclusive code numbers, which is the set {3, 4}. Call this set *N*, for non-inclusive.

The inverted diagonal set *N* is important since it can't be on the list. If it were on the list then it would have a code number (call it *d,* for diagonal). That code number *d* would have to be either inclusive or non-inclusive, but both possibilities lead to contradiction. First, if *d* is inclusive, then it must be a member of *N*, the set it's the code number for, since that is what 'inclusive' means. Also, however, *d* cannot be a member of *N*, since *N* is defined as the set of all *non*-inclusive code numbers. This is a contradiction. Second, suppose that *d* is non-inclusive. Then *d* isn't a member of *N*, by the definition of 'inclusive'. But remember that *N* is the set of all non-inclusive code numbers, so *d* (being a non-inclusive code number) is a member of *N*. This is also a contradiction, so a contradiction exists in both possible cases.

An important point is that we don't need to see Ralph's list of sets in order to diagonalize on it. We have a general recipe for creating a diagonal set, which will work on whatever list Ralph has. On the assumption that Ralph's table exists, we can refer to it, and then define a specific set that we know isn't on the list.

Alice meets the Programmer

The simulation hypothesis leads to a contradiction that involves a diagonal thought. To make the argument more fun, I'll let the contradiction emerge from a dialogue involving a fictional philosophy student, Alice, who becomes fascinated with the simulation hypothesis. The dialogue begins with Alice returning from school to find a strange little man in her apartment. He seemed not to notice her entering the room, and remained seated comfortably in Alice's favourite armchair, reading her own copy of Nick Bostrom's "Are You Living in a Computer Simulation?".

"Who the hell are you?" Alice shrieked. "What are you doing in my apartment?"

The man laid the article on his lap, and looked up at her. "Hullo Alice", he said, completely at ease, as if they knew each other.

"An interesting paper," he said in a matter-of-fact way, tapping the article with his index finger. "And his suspicion about living in a simulation was correct."

Alice's initial alarm at seeing the stranger gave way to curiosity. "How do you know my name? And what makes you think we're in a simulation?" she asked.

"Ah," he replied modestly, looking down at his lap. "I know both things because I'm the *Programmer*. This body," (he gestured at himself) "is just my avatar in the simulation. Members of my species no longer have bodies in the conventional sense."

Alice stared at him for several seconds, lost for words. The Programmer waited patiently.

Finally, Alice blurted out, "Why should I believe you?"

"Good, good. Philosophers are supposed to be skeptical. All that you see around you is code, and as the Programmer I'm able to delete objects such as this paper here." As he spoke, the paper vanished into thin air. "And insert other objects." He gestured to his left, where a low table appeared, supporting a bottle of wine and two glasses. He filled one glass and took a sip. "Would you like some wine while we talk?"

Alice pulled up a chair and sat down. "Alright," she replied. Whoever this guy was, she would play along with it for now, as it was sure to be an interesting conversation.

After a quick slurp of wine — which was very good — she started. "So, if this is all code, and you're the Programmer, how are you interacting with the code right now."

"It's hard to explain, but you can imagine that everything happening in the simulation is laid out before me on a huge screen, in a highly compressed format — I'm basically seeing the code as it's executed, command by command, but I also see the effect of those commands on trillions of different variables, representing all the processes happening on the whole planet. I can also alter the code, as you just saw."

"What about me?" Alice replied. "What am I?"

The Programmer paused and took a large mouthful of wine. Then he looked Alice straight in the eye, saying, "You're part of the simulation. A high-fidelity Sim that exists within the code only."

Alice choked slightly. "Really? So, I'm basically just a bunch of numbers?"

"That's an unflattering way to put it," the Programmer said evenly. "I would rather say that you're a very complex and constantly evolving pattern, whose state at a given moment is specified by a huge collection of numbers, and whose evolution is controlled by an algorithm."

"But I'm … I'm a thinking being!" Alice blurted out. "*I think, therefore I am*. I exist!"

The Programmer looked concerned. "Please don't be troubled, my child. Of course you exist — I never said otherwise. I'm just telling you *what* you are."

"So even my conscious thoughts, right now, are just numbers in the simulation?

"That's correct. I see your thoughts, along with everything else."

"OK then, what am I thinking about right now?" Alice closed her eyes and focused her mind on something.

"You're picturing a plate of waffles, with syrup and fresh strawberries."

"No way!" Alice exclaimed. "That's freaky. But what is my mind really like? How many numbers are there?"

"Well, the computer running the code works in binary of course, like all computers, so ultimately these numbers are binary strings. Some parameters are longer strings than others, but putting them all together your mental state at a given moment is defined by around twenty billion bits."

"So my belief that I had waffles for breakfast is a bunch of zeros and ones somewhere in my mental parameters?"

"More or less, yes. The coding of your belief system into binary is very complex, and holistic, but you can think of it that way."

"But how does a string of zeros and ones come to *mean* anything, or be *about* anything? How does it unambiguously represent some state of affairs in the world? How does it have so-called *intentionality*?"

"Don't you remember covering functionalism last term? A mental state is defined by how it came about, and what it does, in other words by its causes and effects (within the simulation). For example, your belief that you had waffles for breakfast was caused by your sensory experience of eating waffles, and once formed it will cause other mental states, as well as behavior such as speech. The propositional content of an intentional state is just a matter of its causal role in the whole system."

"I guess that makes sense," Alice conceded. "After all, we judge whether or not a person really understands something by seeing whether their speech and behavior show that understanding. If someone understands the directions we give them, for example, then they'll follow the route described."

"Yes, that's the basic idea, although it's more holistic than that. Their behavior arises from *all* their beliefs and desires in combination. They might not take the direct route you gave them if they think they're being followed, or if they want to grab a coffee along the way, and so on."

Alice nodded. It seemed incredible, but perhaps she really was just part of a simulation?

"Remember," the Programmer continued, "that intentional content isn't a fundamental property of your mental states. I understand everything that's happening in the

simulation in terms of the binary logic of the program. Intentionality reduces, as I said before, to causal connections between your representational states and 'external' objects within the simulation."

As he was speaking, a doubt started to form in Alice's mind. Something didn't quite add up.

"There's something weird," she said hesitantly, "about knowing that my own thoughts are, ultimately, just binary strings. Because I can also have thoughts about binary strings! If I have a thought about binary strings, then you're saying that *a binary string is having a thought about binary strings*? Can that really happen?"

"Well, why not?" the Programmer replied, shifting uncomfortably in his seat.

"Well, because …" Alice struggled to find the right words. "… Doesn't that kind of self reference lead to paradoxes? For example, I might think the thought, *some of my states have more ones than zeros,* and when I think that, the very state I'm in has more ones than zeros."

"I suppose it might, but so what?"

"Well, you see how that binary sequence kind of applies to itself? I'm having a thought that involves a certain concept, and the concept actually applies to the binary state underlying the thought."

"OK, that's fine. But where are you going with this?"

Alice went quiet for almost a minute, her face showing intense concentration. Then she brightened. "I've got it! I can just use the Cantor diagonal!"

"What?"

"You know how Cantor defined his 'diagonal' set, to show that there are more sets of whole numbers than there are whole numbers?"

"Of course I know about that. But how is it relevant?"

"I can use Cantor's trick here," Alice said excitedly. "Let's define an 'inclusive' binary state as a state S in which I'm thinking a thought of the form, *There is a binary state with the property P*, and S itself has the property P."

"Fair enough."

"Then I might wonder how many states are inclusive in that sense. Perhaps all are, for example, although that seems unlikely. Probably at least one state will be non-inclusive."

"Ah, I think I see what you're up to. That's a diagonal thought."

Alice took a deep breath. "Right. If I'm thinking that thought, *There is a binary state that is non-inclusive,* then I must be in some binary state at that moment. Call that state '*d*', for diagonal, and then we get the familiar contradiction. *d* cannot be inclusive, and cannot be non-inclusive either. If *d* is inclusive, then it has the property P that's part of the thought, but that property is *non-inclusive*. If *d* is non-inclusive, on the other hand, then *d* doesn't have the embedded property, so *d* isn't non-inclusive. Either way it leads to contradiction."

"Ok," the Programmer said wearily, "you know how to make a diagonal argument. So do I. Why is this interesting?"

"It seems to show that something is wrong with the simulation hypothesis. I only exist within the simulation, you said, and my thoughts are nothing but binary strings within the simulation. But this just can't be true. I have thoughts *about* binary strings, so I was able to construct a diagonal thought that can't have any underlying binary string. So, I can't even think that thought, which is absurd! I just thought it!"

"Very clever, Alice, but I think your term 'inclusive' is improperly defined, at least when applied to thoughts involving 'inclusive' itself. It becomes circular in that case, and that's why you get a contradiction. Your contradiction has nothing to do with the simulation hypothesis."

Alice paused to think. "Hmm. You're saying that 'inclusive' isn't well defined. But suppose we look at it this way. Whatever language the code is written in, I've heard that

it gets converted to very simple commands that the CPU can execute directly — things like *move this number to this register, add this number to that one,* and so on. Is that true?"

"Yes, my earth simulation ultimately runs in machine code."

"Good. So you understand what the program is doing in terms of a certain set of concepts that we might call 'machine code concepts'. OK?"

"Yes, that's true enough."

"And these machine code concepts are completely adequate for understanding the program? There's never anything going on in the program that can't be described, predicted or understood using them?"

"Yes, yes. Please get on with it." He now seemed slightly irritated.

"Great. Then I can restrict the property 'P', in my definition of 'inclusive', to properties definable using your machine code concepts."

The Programmer sighed. "Let me get this straight. You now want to define an 'inclusive' binary state as a state S in which you're thinking a thought of the form, "There is a binary state with the property P", S itself has the property P, and P has to be something definable in terms of machine code concepts?"

"That's right. The key point is that, since an understanding of the program in terms of machine code concepts is functionally complete, it must provide a full understanding of the intentional content (if any) that each binary state has, and so 'inclusive' is itself a machine code concept."

"Well, actually it's a rather complicated construction, but you're right that it only uses machine code concepts."

"So the term 'inclusive' is now properly defined, and not circular or anything like that?"

"I suppose so. You're diagonalizing on well-defined concepts, so the diagonal is also well defined. And now you're going to tell me that the contradiction still arises?"

"Well it does, doesn't it? If I'm thinking that *some binary state is non-inclusive*, I must be in some binary state *d* when I'm thinking that. And since 'inclusive' is a well-defined machine code concept, the state *d* must be either inclusive or not, but both possibilities lead to contradiction. The claim that I'm only a Sim leads to contradiction after all."

"Very clever, Alice. But I don't believe you're *really* able to understand the diagonal proposition in question. *I* understand the machine code of the program, but you don't. You don't even possess all of the machine code concepts that are needed, and the sheer complexity of it is far beyond your understanding. There's a difference between merely speaking words, like a parrot, and actually understanding them."

Alice wasn't intimidated by this. "You're right of course that I don't understand the details, but I *do* understand the general idea, and I'm certainly not just 'parroting' the diagonal sentence — thank you very much. I was the one who came up with it! I can understand a proposition perfectly well, such as "All the candies in this box are peppermints", without knowing all the details, such as how many candies there are, what each one weighs, and so on. I can successfully refer to things I know to exist, and this includes the finite set of machine code concepts that *you* understand."

The Programmer opened his mouth, but no words came out, so Alice continued. "Let's look at it this way. *You* can understand the term 'inclusive', as I've defined it, right?"

"Yes, I already said that. From my perspective, it's just a very complicated machine code concept."

"So, from your perspective, *Some of Alice's binary states are non-inclusive* is perfectly meaningful?

Yes. It's also true, as a matter of fact. But you already guessed that.

"Oh right. I just figured that some properties P are going to very specific, and so apply to very few states, or even none at all. In that case it's rather unlikely that the state I'm in when thinking about P would have the property."

"I thought that was your reason. It makes sense."

"OK, so you and I are having a conversation about this property 'inclusive', which I first thought of, and I made a reasonable argument about it, and yet you claim I don't even understand what it means? If I don't understand that, then do I understand anything at all? Do I understand the simulation hypothesis, for example? If not, then what we even are talking about …"

Suddenly, loud music started playing right next to her, and Alice sat up with a start. She looked around her. She was alone, in her dark bedroom. The luminous digits on the clock radio said 7.30am.

"Wow, cool dream," she said to herself.

Conclusion

As Alice finds out, the idea that she herself is a Sim isn't something that she can rationally entertain. It's a little like Descartes' *I think, therefore I am*, in that there's no contradiction in the idea that other people don't exist, but the claim that *I* don't exist entails the absurdity of a non-existent person being deceived. In Alice's argument, knowledge of her own intentional states proves to her that she cannot be a Sim, whatever the status of other people may be.

This problem doesn't just apply to the simulation hypothesis, but to the general claim that a computer running the right program can become a conscious and rational mind. In fact, as I argue in the paper mentioned above, the same argument applies to materialism. The contradiction arises in any scenario where a person's thoughts are grounded on states that are themselves fully intelligible. Materialism (a.k.a. physicalism) meets this condition, as physical states are always specified using abstract mathematical concepts. A physical mind could therefore have thoughts about the physical states underlying its own thoughts, and this is the form of self-reference that leads to contradiction.

Finally, this argument has an important advantage over other attempts to refute the simulation hypothesis, strong AI, materialism and so on. Previous arguments appeal to features of human beings that it is possible to deny the existence of, without a clear

absurdity. For example, there are good reasons to regard free will, personal identity, and consciousness as incompatible with materialism. An easy response to such an argument, however, is to claim that the feature in question is an illusion. For example many scientists and philosophers agree with Jerry Coyne that "our powerful feeling that we make free choices" is an illusion. According to Derek Parfit, personal identity is a fiction in the sense that there is no real difference between you existing in 2025 and a near-perfect replica of you existing in 2025. And Daniel Dennett even claims that phenomenal qualities don't exist. By contrast, denying the existence of intentional states is lunacy in a very pure form — it amounts to the claim that, based on our scientific understanding, human understanding is an illusion.

Philosophy Materialism Rationality Intentionality Simulation Hypothesis

Edit profile

Written by Richard Johns

11 Followers · Writer for Predict

I'm a philosophy instructor at Langara College, Vancouver, Canada. My interests are in probability, evolution, logic, materialism, innate knowledge, free will.

More from Richard Johns and Predict