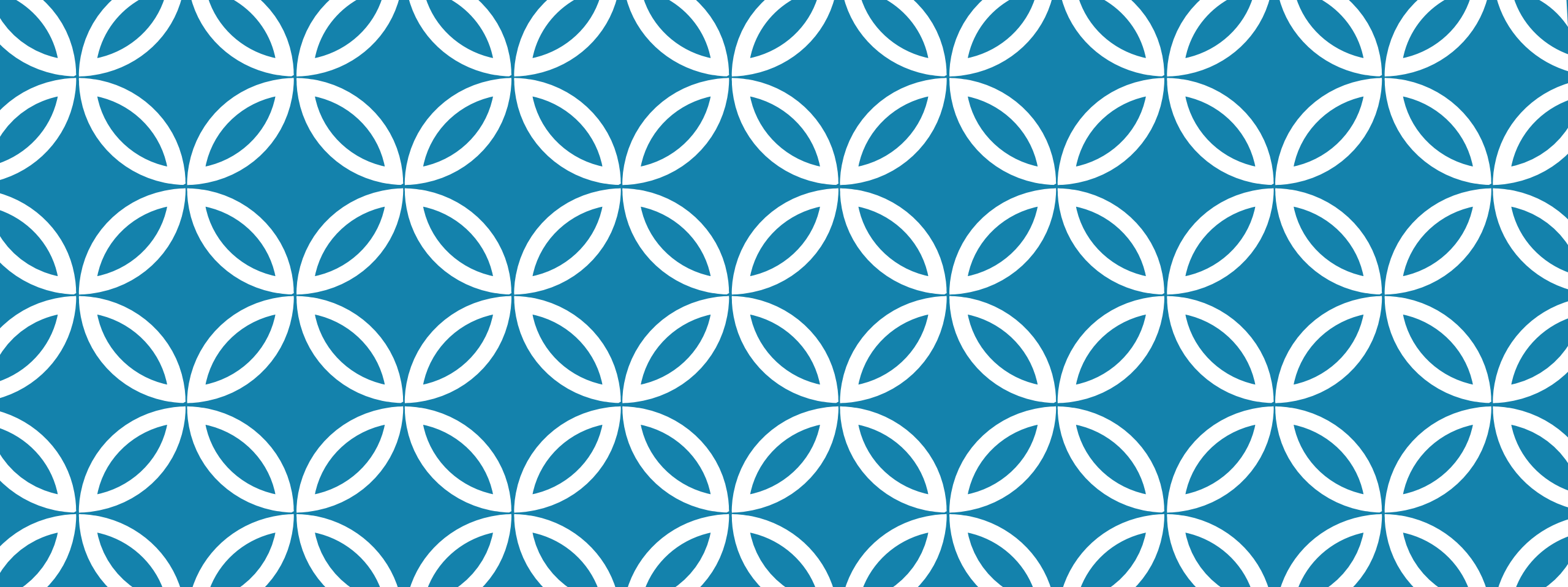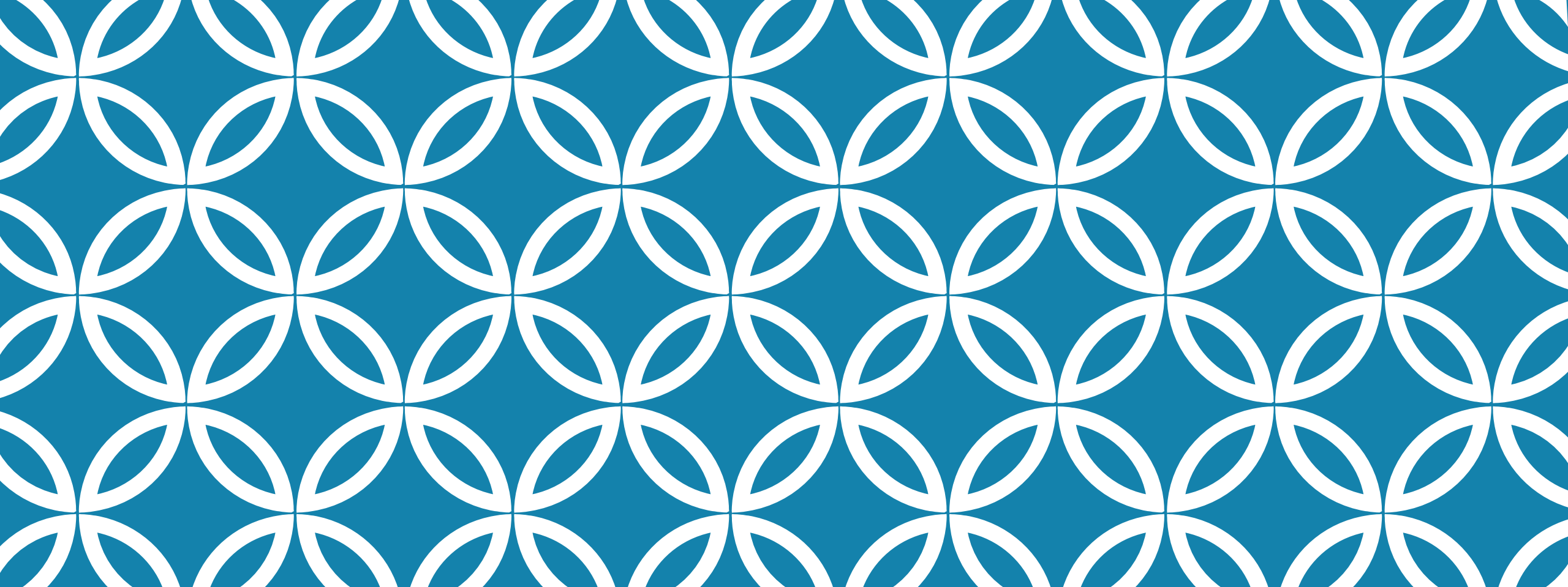# WHY BRAINS IN VATS WOULD NOT BE (MUCH) DECEIVED

Benoit Gaultier
(University of Zurich)
Collège de France 05/06/23

WHY BRAINS IN VATS WOULD ~~NOT~~ BE (MUCH) DECEIVED

Benoit Gaultier
(University of Zurich)
Collège de France 05/06/23

WHY BRAINS IN VATS WOULD NOT BE (MUCH) DECEIVED AT ALL

Benoit Gaultier
(University of Zurich)
Collège de France 05/06/23

AIM OF THIS TALK: DISCUSSING CHALMERS' VERIDICALISM

More precisely:

i) Indicating **a reason to doubt** the cogency of **one of Chalmers' arguments for veridicalism.** [slides 12–28]

ii) Suggesting that **if BIVs' ordinary beliefs are true for the reasons indicated by Chalmers, then BIVs' (metaphysical) beliefs about the fundamental nature of things or reality also are true.** [slides 29–35]

iii) Suggesting that it is compossible that **BIVs' beliefs about the fundamental nature of reality are true,** and that **BIVs are ignorant of the fundamental nature of reality.** [slides 36–37]

CHALMERS' VERIDICALISM: REFERENCES

"**The Matrix as Metaphysics**" (in Christopher Grau, ed., *Philosophers Explore the Matrix* (OUP, 2005), 132–76) = **C1**.

"**Structuralism as a Response to Skepticism**" (*Journal of Philosophy* 115 (2018): 625–60) = **C2**.

Reality+. Virtual Worlds and the Problems of Philosophy (W. W Norton & Company, 2022) = **C3**.

THE TYPE OF BIV SCENARIO I SHALL FOCUS ON

"We could distinguish between **temporary and permanent** simulations (Do people enter the simulation for a brief period or spend their whole lives there?), **perfect and imperfect** simulations (Do we faithfully simulate all the laws of physics, or do we allow approximations and exceptions?), and **pre-programmed and open-ended** simulations (Is there a single course of events programmed in advance, or can various things happen depending on initial conditions and what the sims choose?)." (C3)

"Here's another distinction. The **global**-simulation hypothesis says that the simulation simulates a whole universe in detail. For example, a global simulation of our universe will simulate me, you, everyone on Earth, planet Earth itself, the whole solar system, the galaxy, and everything beyond. The **local-simulation** hypothesis says that the simulation simulates only a part of the universe in detail. It might simulate just me, or just New York, or just Earth and everyone on it, or just the Milky Way galaxy." (C3)

We **are and always have been** in a **global** and **open-ended** artificially designed computer simulation that "precisely mirrors" – or **perfectly** simulates – the world or universe it is simulating and the physical laws it obeys.

HOW THE SIMULATION WORKS

In a non-simulated world W1, "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 [in W1]. To do this, the brain is connected to a giant [perfect, global and open-ended] computer simulation of [W1]. 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 [in W1] it lacks a body." (C1)

The envatted brain is "associated with a particular simulated body" in the following way: "whenever this [simulated] body receives [simulated] sensory inputs inside the [simulated world W2], the envatted cognitive system will receive [non-simulated] sensory inputs of the same sort. When the envatted cognitive system produces motor outputs, corresponding outputs will be fed to the [simulated] motor organs of the simulated body". (C1)

As a result, my belief that I'm currently giving a talk at the Collège de France, that I have two hands, that I'm French, etc. is caused/underlaid by a computer simulation.

HOW THE SIMULATION WORKS

So stated, the simulation hypothesis in question is neutral

i) on the question of where I am located, and hence on whether I am a BIV physically connected to a computer simulation [= located in W1], or a simulated person whose name in the simulation is Benoit [= located in W2], or both.

ii) on the question of what my different beliefs refer to.

ON THE NATURE OF SIMULATED ENTITIES

Simulated entities are "digital entities, **made of** computational and informational processes. More succinctly, they're **made of bits**. They… are **grounded in a pattern of bits** [or "structures of binary information"] in a computer." (C3)

"The bits here are physical bits, embodied in voltages in integrated circuits or some other physical basis". (C3)

A simulated "chair or table is **made of digital processes, just as a physical chair or table is made of atoms and quarks and ultimately of quantum processes**". (C3)

Simulated "tables are **constituted by** computational processes" (and facts about such tables "supervene on computational facts") while "tables are constituted by quantum processes" (and facts about such tables "supervene on quantum-mechanical facts"). (C1)

ON THE TRUTH-VALUE OF OUR BELIEFS IN THE BIV SCENARIO IN QUESTION

This version of the simulation hypothesis does not imply that we are "massively deluded", that we "have massively false beliefs about the external world". On the contrary, they are "largely correct" partly because the things they are about – cats, chairs, trees, cars, Paris, Beyoncé, etc. – are all real. "They're just made of bits" (C3) or "constituted by patterns of bits" (C1).

"If I grew up inside Sim Universe, then I've been applying the word 'tree' to digital trees my whole life. Digital trees are what I mean by 'tree'. If I grew up outside all simulations, then I've been applying the word 'tree' to nondigital trees my whole life. Nondigital trees are what I mean by 'tree'. So I'll describe Sim Universe differently depending on whether I've grown up inside it or not. If I've grown up in Sim Universe, I'll say it contains trees, because 'tree' for me means 'digital tree'. If I've grown up outside all simulations, I'll say that Sim Universe doesn't contain trees, since 'tree' for me means 'nondigital tree'." (C3)

ON THE TRUTH-VALUE OF OUR BELIEFS IN THE BIV SCENARIO IN QUESTION

So, in the scenario in question, "most of our ordinary beliefs will be true" despite the fact that "the trees and cars and Beyoncé are not exactly how we [think they are]": they are not "ultimately made of fundamental particles such as atoms and quarks; instead, they're made of bits". (C3)

= "their fundamental reality [or "underlying nature"] is a little different from what we [think]". (C3) But this does not prevent most of our beliefs about them to be true.

The fact that these things "are realized computationally in a way that we might not have originally imagined […] does not contradict any of our ordinary beliefs. At most, it will contradict a few of our more abstract metaphysical beliefs". (C1)

If we believe that "the level of quarks in our universe [is] the bottom level of reality", that our "universe is the ultimate reality", that we're not in a simulation, "that flowers are not digital", these beliefs are wrong. But they are "mostly scientific or philosophical beliefs about reality. Undermining them does not undermine everyday beliefs such as 'There are flowers blooming in the garden'", that will still be true.

WHY BEING WRONG IN THE WAY WE ARE IN THE BIV SCENARIO DOES NOT MAKE OUR ORDINARY BELIEFS FALSE?

Let's admit that, if the BIV scenario in question obtains, then

i) (given certain structuralist considerations) the ordinary beliefs I express by saying "There are flowers blooming in the garden", or "There are chairs in the room" are about digital flowers and seats;

Ii) digital flowers and chairs are as real as nondigital chairs and flowers – it's just that the former are made of bits, contrary to the latter.

What exactly supports the claim that being wrong in the way I am in the BIV scenario about the fundamental or underlying nature of the things my ordinary beliefs are about does not make them false?

This claim is not supported by structuralist considerations, but by "first-order considerations about various scenarios". (C2)

A FIRST COMPARISON

If the BIV scenario obtains, "physics as we know it is 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, 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." (C1)

"Science has taught us that there's much more to reality than initially seems to be the case. For millennia, we didn't know that cats and dogs and trees are made of cells, let alone that the cells are made of atoms or that those are fundamentally quantum mechanical. Yet these discoveries about the nature of cats and dogs and trees have not undermined their reality. If I'm right, the discovery that we're in a simulation should be treated the same way. It will be a discovery about the underlying nature of cats and dogs and trees—that they spring from digital processes—but it won't undermine their reality." (C3)

A FIRST COMPARISON

"The ancients thought that trees were compounds of earth, air, fire, and water. Nineteenth century scientists thought they were clouds of atoms balanced in mutual equilibrium. Most physicists today think they are quantum states defined on a high-dimension Hilbert space. Despite these disagreements about the underlying nature of trees, the ancients, Victorians, and contemporary physicists agree that there are trees. Neither the transition from ancient physics to classical physics, nor the transition from classical physics to quantum physics, prompted an outbreak of skepticism or antirealism about trees. [...] if we change our minds about the underlying nature of trees – say, by abandoning classical in favor of quantum mechanics – we don't change our minds about the existence of trees. We just change our minds about what ultimately accounts for the fact that there are trees." (Michael Pelczar. "Why Idealism?" (2022))

"Chalmers's stance receives support from the history of science. People's beliefs about the underlying nature of macroscopic physical phenomena have changed dramatically over the centuries, from combinations of the Four Elements, to geometric configurations of Democritean atoms, to dynamical systems of Newtonian bodies, to excitation states of quantum fields. Despite these changes, people's beliefs about the world's macroscopic physical contents have remained highly stable. The ancient Greeks, the natural philosophers of the Enlightenment, and scientists of the 21st century all agree that the world contains rocks, despite having very different beliefs about the underlying nature of rocks." (Michael Pelczar. *Phenomenalism* (2023): OUP)

THE FIRST COMPARISON, SUMMARISED

The successive scientific discoveries that things like trees, chairs, or cats are not fundamentally made of what we thought they were did not lead us (rightfully, intuitively) to think that our ordinary beliefs about them were wrong – e.g. to think that Aristotle's belief that they are trees in the Lyceum's garden was false – but just that our beliefs about their underlying nature were so.

Why should things be different when it comes to discovering that trees, chairs, or cats are not in fact fundamentally made of quarks and quantum processes, contrary to what we thought before this discovery, but made of of structures of bits?

A SECOND COMPARISON

"Suppose God announces that the it-from-bit hypothesis is true [i.e physical processes are real but "there is a digital level underlying physics: roughly speaking, molecules are made of atoms, atoms are made of quarks, and quarks are made of bits"]. Underneath traditional physics is a level of interacting bits […]. Will we then conclude that nothing is real? I don't think so. Discovering atoms didn't make us reject molecules. Discovering quarks didn't make us reject atoms. So discovering bits shouldn't make us reject quarks. If the it-from-bit hypothesis is correct, there are still quarks, cats, and chairs. It's just that the cats and chairs are made of atoms which are made of quarks which are made of bits." (C3)

If, contrary to what we think, a digital level underlies physics in the non-simulated world we live in, and this does not make our ordinary beliefs false, why should things be different when it comes to simulated worlds in which a digital level underlies physics?

A THIRD COMPARISON

Suppose God announces that "at the most fundamental level, [physical processes] are constituted by processes in [her] mind". (C1)

= Suppose that Berkeley's idealism is true and that the ordinary objects we encounter in episodes of perceptual experience, and have beliefs about (e.g. apples, stones, trees, tables, and chairs), are collections of "ideas" caused by God, which "appear in our consciousnesses as the effects of his causal activity" (Grayling 2005, "Berkeley's argument for immaterialism").

This would be "the *metaphysical* way of describing what, in ordinary terminology, we describe as seeing trees, tasting ice cream, and so forth. The latter way of describing the facts is not incorrect" (*art. cit.*).

If, contrary to what we think, ideas underlie physics in the non-simulated world we live in, and this does not make our ordinary beliefs false, why should things be different when it comes to simulated worlds in which ideas underlie physics?

THE THREE COMPARISONS, SUMMARISED

Living in a non-simulated world where we don't know

that quantum theory/ the it-from-bit hypothesis/ the Berkeleyan view is true of the objects our ordinary beliefs are about

does not make these beliefs false.

So, living in a simulated world where we don't know

the fundamental nature of the objects our ordinary beliefs are about – i.e. that these objects are made of bits

does not make these beliefs false.

SMITHSON'S IDEALISM

"Suppose it turns out that, in fact, we are in an envatted brain scenario. Suppose further that we come to learn this somehow […]. Upon learning this result, we might initially react by saying things like: 'Everything we see is merely an illusion!' […] But this shock would pass. And after several minutes, we would return classifying experiences as illusory or veridical just as we did before. We would, for example, return to saying things like: 'The stick's appearance is deceptive; it is actually straight'." (Robert Smithson (2020). "Idealism and illusions". *European JoP*)

More generally, "it seems that our judgments about objects would continue to be correctly assertible even after [learning] what the external world is like 'in itself'." (Robert Smithson (2023). "Edenic Idealism". *Australasian JoP*)

How to explain this?

Our ordinary "distinction between illusory and veridical experiences ultimately [does not] hinge on facts about some external reality that is completely independent of human experience" (though giving rise to it). (Smithson 2020)

METAPHYSICAL AND PHYSICAL HYPOTHESES

The BIV scenario in question is a "metaphysical hypothesis" that is "analogous to a physical hypothesis, such as one involving quantum mechanics. Both [hypotheses] tell us about the processes underlying chairs". (C1)

While physical hypotheses are "concerned with the microscopic processes that underlie macroscopic reality", metaphysical hypotheses are concerned with "the reality that underlies physics itself". (C1)

The BIV scenario in question is the metaphysical hypothesis that "standard physics is grounded in a deeper level of computational physics". (C2)

"In Chalmers's view, discovering that we live in a computer simulation would be like discovering that what underlies physical objects are protons, neutrons, and electrons, rather than Earth, Air, Fire, and Water. It would be a discovery about the metaphysical rather than the physical nature of physical things." (Pelczar 2023)

THE THREE COMPARISONS, SUMMARISED*

Living in a non-simulated world where we don't know

that quantum theory (Sc)/ the it-from-bit hypothesis (Sc/MP?)/ the Berkeleyan view (MP)
is true of the objects our ordinary beliefs are about

does not make these beliefs false.

So, living in a simulated world where we don't know

the fundamental nature of the objects our ordinary beliefs are about – i.e. that these
objects are made of bits (MP)

does not make these beliefs false.

ABOUT TWO (RELATED) DIFFERENCES

[My aim in slides 23–28: rejecting the foregoing inference = arguing that there is a difference between the BIV scenario and the three other scenarios that makes this inference problematic/disputable.]

It cannot be argued that what makes the BIV scenario relevantly different from the three other scenarios is that the (level of) reality R_{n-1} that underlies the (level of) reality R_n (whether macroscopic or microscopic) in the BIV scenario does not determine R_n in such a way (i.e. causally ?) that R_{n-1} would be empirically discoverable.

Because this is also the case in the Berkeleyan scenario (and in the it-from-bit scenario in a non-simulated world?).

The same goes for the idea that a relevant difference between the BIV scenario and the three other scenarios is that the (level of) reality R_{n-1} that underlies the (level of) reality R_n (whether macroscopic or microscopic) in the BIV scenario does not explain observable phenomena in R_n in the way in which science does.

= This is also the case in the Berkeleyan scenario (and in the it-from-bit scenario in a non-simulated world?).

ANOTHER DIFFERENCE

What makes the BIV scenario much different from the three other scenarios is that, in the BIV scenario, i) what underlies the simulated world is not part of it (and ii) we do not have any cognitive access to the underlying reality in question).

= Being wrong about what underlies the simulated world is not

being wrong about certain things that are true **in** this world (= what we are wrong about in the three other scenarios), but

being wrong about certain things that are true **of** this world – e.g. that this world has a supervenience base located outside of it/ that this world is underlaid by certain realities located outside of it).

Maybe being wrong about these things is sufficient for making most of our beliefs wrong in the BIV scenario.

But maybe it is also necessary that we do not have any cognitive access to the underlying reality in question. Why this additional condition? Because, in the following thought experiment, (i) is satisfied, but despite this I tend to have the intuition that the BIVs' ordinary beliefs are true.

A FIFTH SCENARIO: THE NIGHT BRAIN STIMULATION SCENARIO (1/2)

Consider the following variant of our BIV scenario, where the simulation is also perfect, global and open-ended, but only temporary.

Unbeknown to me, during the night that follows my 25th birthday, scientists take control of my brain for a few hours, when I'm asleep. They stimulate my brain in such a way that, while my body remains inert in my bed, the inputs my brain receives from them are such that I (re)experience the world exactly as I did during my birthday before falling asleep (without being aware of this). Well, not exactly… Whereas, in the evening, I succeeded *in extremis* in catching a train I had to take, the scientists stimulate my brain during the night in such a way that I experience missing the train *in extremis*, and what would then have happened during the rest of the day (given the laws of nature and supposing the truth of determinism).

TRAIN I

TRAIN II

A FIFTH SCENARIO: THE NIGHT BRAIN STIMULATION SCENARIO (2/2)

On the next night, they stimulate my brain in such a way that I experience everything I would have experienced during the day if I had missed my train the day before in the way I experienced missing it. They go on like this on every night.

When I wake up in the morning, the state of my brain is exactly as it would have been had the scientists did not stimulate my brain during the night. There is no neural trace of the night brain stimulation. I have absolutely no memory of it, and it does not affect or determine in any way how my life goes – my life will be until my death exactly as it would have been without these night interventions.

ON THE TRUTH-VALUE OF MY BELIEFS IN THE NIGHT BRAIN STIMULATION SCENARIO

It seems like on every night I live a second or alternative life – the life I would have lived, in the world in which scientists take control of my brain during the night, if I had missed my train in the way in question.

I don't have the intuition that the beliefs I would form during the night would be false – even when it comes to the beliefs I would form during the very first night that would contradict the beliefs formed during the evening, after succeeding in catching the train.

(Interestingly(?), I have on then contrary the intuition that the ordinary beliefs of recently envatted people would be false.)

DOES CHALMERS' VERIDICALISM IMPLY THAT BIV'S BELIEFS ABOUT THE FUNDAMENTAL NATURE OF THINGS ALSO ARE TRUE?

According to Chalmers' veridicalism, if we are in the BIV scenario, my belief that there are chairs in the garden is true.

The same goes for:

i) my belief that these chairs are made of grey wrought iron (and not of e.g. while plastic).

ii) my belief that these chairs then are made of something whose molecular structure is XYZ (that is distinctive of iron).

iii) my belief that these chairs then are made of the quarks and quantum processes that are involved in such a molecular structure.

DOES CHALMERS' VERIDICALISM IMPLY THAT BIV'S BELIEFS ABOUT THE FUNDAMENTAL NATURE OF THINGS ALSO ARE TRUE?

What about the belief that these chairs are not made of something more fundamental? Does the obtaining of the BIV scenario make this belief false?

The chairs can't be said to be made of bits in the sense of "made of" in which my belief that the chairs are made of quarks and quantum processes is true. Because:

i) Things T cannot be said to be made of things T* that are not part of the world to which T belong.

ii) The concerned bits are not part of the world the chairs are part of.

So, we would not be wrong to believe that the chairs are not made of bits in the appropriate sense of "made of" in which we are right to believe that the chairs are made of quarks and quantum processes.

DOES CHALMERS' VERIDICALISM IMPLY THAT BIV'S BELIEFS ABOUT THE FUNDAMENTAL NATURE OF THINGS ALSO ARE TRUE?

But what about the belief that these chairs are not "**underlaid by**", or "**grounded in**", patterns/structures of bits? Is this belief false?

If structuralism is true, IS A PATTERN OF BITS could refer either to

i) whatever plays the is-a-pattern-of-bits-role (/to certain observable behavioral dispositions), "without any regard for whether the realizers of [this] role are the same as that in the actual world" (as Grace Helton summarises it), or to

ii) whatever realizes this role (/these dispositions) in some world considered as actual.

DOES CHALMERS' VERIDICALISM IMPLY THAT BIV'S BELIEFS ABOUT THE FUNDAMENTAL NATURE OF THINGS ALSO ARE TRUE?

Under the first option, when we believe or say that the chairs in the garden are not underlaid by patterns of bits, we refer to what plays the is-a-pattern-of-bits-role in our simulated world – which would make it **true** to believe, about the chairs we refer to when we rightly believe that there are chairs in the garden, that they are not underlaid by patterns of bits.

In other words, under the first option, if we were in a simulated world, the chairs underlaid by patterns of bits would just be the chairs we see on e.g. computer screens when playing videogames.

DOES CHALMERS' VERIDICALISM IMPLY THAT BIV'S BELIEFS ABOUT THE FUNDAMENTAL NATURE OF THINGS ALSO ARE TRUE?

Under the second option (which seems to be correct regarding IS A PATTERN OF BITS and natural kind terms), when we believe that the chairs in the garden are not underlaid by patterns of bits, we refer to whatever realises the is-a-pattern-of-bits-role in some world considered as actual (i.e. the realisers of this role in that world).

Let's try to see what this second option implies through two questions:

1) What is the actual world for us when we believe that the chairs in the garden are not underlaid by pattern of bits?

If the answer is "the simulated world in which the chairs are not underlaid by pattern of bits", then when we believe that the chairs in the garden are not underlaid by patterns of bits, this belief is true.

DOES CHALMERS' VERIDICALISM IMPLY THAT BIV'S BELIEFS ABOUT THE FUNDAMENTAL NATURE OF THINGS ALSO ARE TRUE?

2) What realises the is-gold-role when we rightly believe, in the BIV scenario, that there is gold in the mine?

Suppose we live in a non-simulated world. What realises the is-gold-role in this world?

A) A certain distinctive microstructure XYZ ?

B) The types of quarks and quantum processes actually involved in this microstructure?

Ordinarily, we would say that what makes it the case that a certain substance is gold is that the microstructure of this substance is XYZ. This *maybe* supports the idea that what realizes the is-gold-role in a non-simulated world – *especially if* this microstructure is multiply realisable (e.g. if it is possible that other types of quarks realise it).

But if what realises the is-gold-role is determined at this level in a non-simulated world, why should things be different when it comes to what realizes the is-a-pattern-of-bits-role in a simulated world?

And if they are not different, and if by IS A PATTERN OF BITS we refer to the realizers of the is-a-pattern-of-bits-role in some world considered as actual, then when, in simulated world, we believe that the chairs in the garden are not underlaid by patterns of bits, this belief is true, as it is true at the non-fundamental microscopic level in question.

DOES CHALMERS' VERIDICALISM IMPLY THAT BIV'S BELIEFS ABOUT THE FUNDAMENTAL NATURE OF THINGS ALSO ARE TRUE?

Then, under both options, if the BIV scenario obtains, our belief or claim that the chairs we refer to when we rightly believe that there are chairs in the garden are not underlaid by patterns of bits is true.

The moral is, if (the way in which Chalmers argues for) veridicalism is correct, then our beliefs about the fundamental nature of things also are true.

WHY BIVS CAN BE RIGHT ABOUT THE FUNDAMENTAL NATURE OF THINGS WHILE BEING IGNORANT OF IT

Even if Chalmers' veridicalism implies that, in the BIV scenario in question, even our beliefs about the fundamental nature of things are true, and hence that we are not deceived *at all* in this scenario, this should not be taken as a *reductio ad absurdum* of these arguments.

Because the intuitive desideratum to be satisfied is that if the BIV scenario obtains, we are (doomed to be) ignorant of certain things.

And being (doomed to be) ignorant of certain things does not imply being wrong about them.

WHY BIVS CAN BE RIGHT ABOUT THE FUNDAMENTAL NATURE OF THINGS WHILE BEING IGNORANT OF IT

Indeed, it is possible for us to be both

i) ignorant of the fact that the BIV scenario obtains, and that the things we refer to in our ordinary beliefs are underlaid by (physical) realities that are not part of the world these things are part of

ii) right when believing this is not the case, due to the fact that we cannot refer to such realities.

So, what we could be afraid of

is not of being wrong on whether the things we refer to in our ordinary beliefs supervene on (physical) realities that are not part of the world these things are part of;

but of being incapable of referring to these realities that, if they obtained, would make people who can refer to them right to believe that the things we refer to in our ordinary beliefs are underlaid by (physical) realities that are not part of the world these things are part of.

LA RÉUNION DE L'ANGOISSE
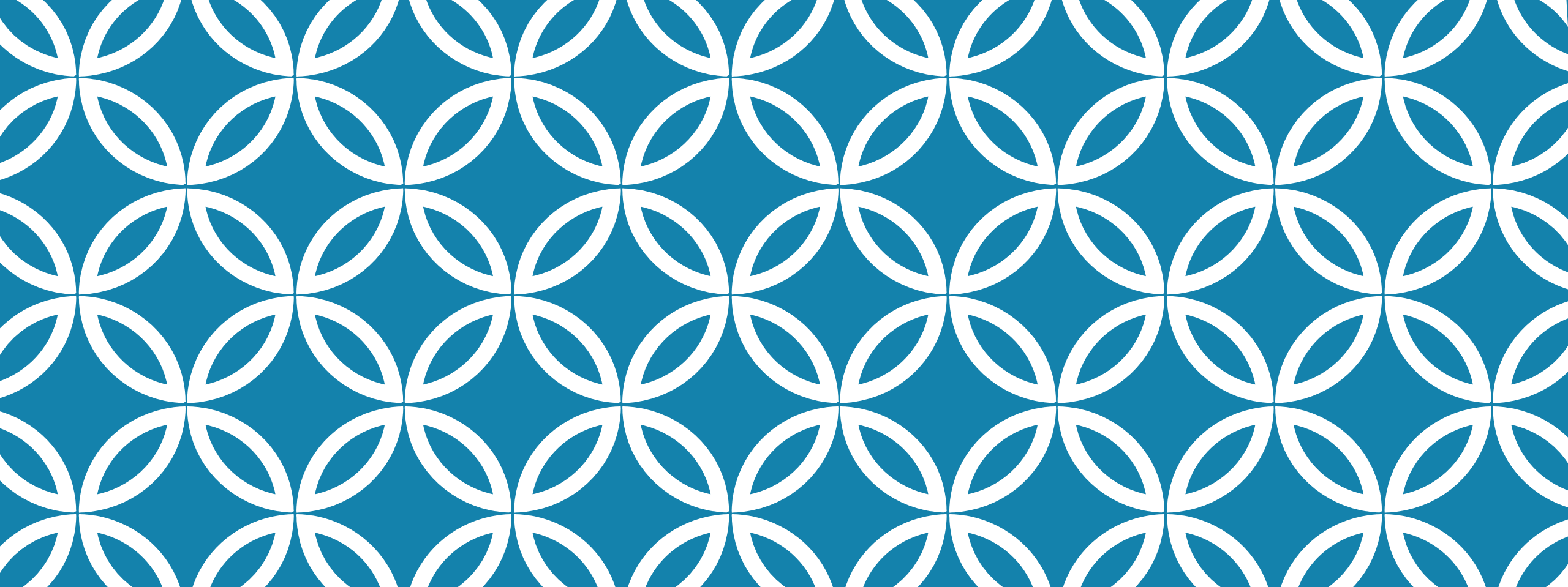
THANK YOU!

For any question:
benoit.gaultier@philos.uzh.ch