PRINCIPIA **27**(2): 263–279 (2023) doi: 10.5007/1808-1711.2023.e88067
Published by NEL — Epistemology and Logic Research Group, Federal University of Santa Catarina (UFSC), Brazil.

A SCRUTINY OF THE TWO-DIMENSIONAL ARGUMENT AGAINST PHYSICALISM

WILSON MENDONÇA
Universidade Federal do Rio de Janeiro, BRAZIL
mendonca@ifcs.ufrj.br

JULIA TELLES DE MENEZES
Universidade Federal Fluminense, BRAZIL
juliatelles@gmail.com

Abstract. Chalmers's two-dimensional argument against materialism (aka the zombie argument) is arguably the most ingenious attempt to ground a view about fundamental reality on epistemic considerations. From the conceivability of a being that is physically identical to a conscious being but that is deprived of phenomenal consciousness (a zombie), the argument draws on the interplay of the primary and the second intensions of the zombie hypothesis to infer the metaphysical possibility of a zombie world, and thus the falsity of physicalism about phenomenality. By means of a detailed reconstruction of the two-dimensional argument, the paper tries to isolate its most central assumption: that the role played by an epistemic scenario (an intentional object) in the verification of the zombie hypothesis is played by a non-intentionally individuated metaphysical world (the zombie world) considered as actual. The paper argues that no non-viciously circular case for this assumption has been made. Thus, the two-dimensional argument is at best inconclusive.

Keywords: epistemic possibilities • intensions • metaphysical plenitude • metaphysical possibilities • strong necessities

RECEIVED: 21/05/2022 REVISED: 19/01/2023 ACCEPTED: 18/02/2023

1. Epistemic arguments

When it comes to oppose the claims of physicalism in philosophy of consciousness, epistemic arguments loom large. The key premises in these arguments concern how things are with us as cognitive, conscious agents: what we authoritatively know about the internal world, what we clearly perceive by way of introspection, what we can consistently conceive or imagine about our existence in the broadly physical world. The conclusion is unabashedly metaphysical. It concerns objective reality. It says that physicalism, the doctrine according to which everything that exists in reality is identical or reducible to the physical, is false: at least conscious states are not physical or not reducible to the physical.

© 2023 The author(s). Open access under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License.

This argumentative trend was arguably inaugurated by Descartes. The operative starting point in Descartes is the idea which he takes to be as clear and distinct as any idea can ever be, namely, that I could possibly think and feel even if I did not have a body. This should show that any actual correlation between conscious episodes, on the one hand, and bodily events, on the other, is at best only a contingent fact. Conscious events are not necessarily accompanied by physical events, for I can clearly form a conception of a situation in which the correlation fails. The fact that the conscious mind and the body may be conceived as distinct entities is a tool to infer that they could indeed be distinct, and assuming with Kripke (1971) that identities are always necessary, the possibility of their being distinct is then used to conclude that they are indeed distinct. Therefore, physicalism is false.

In a more contemporary frame of mind, Joseph Levine (1983) explores the intuition that a putative psycho-physical identification like *pain = firing of C-fibers* would leave as unexplained the very nature of pain. The intuitively compelling claim that there is an explanatory gap in the psycho-physical case, but not in the physical-physical identification *water* = H_2O, for instance, then grounds the rejection of physicalism about the conscious mind.

Last but not least, let us consider Frank Jackson's knowledge argument (Jackson 1982). It is famously based on the strong intuition to the effect that Mary, the brilliant scientist who learned everything there is to be learned about the physics and the physiology involved in the perception of colors, acquires new knowledge when she leaves the black and white room where she spent her whole live and finally sees a red object. The newly acquired knowledge of what it is like to see red cannot be the apprehension of a physical fact, for by hypothesis all physical facts have been apprehended by Mary while she was in the black and white room. Thus, contrary to physicalism, knowable reality contains more than physical entities.

What we have so far is a hopefully representative sample of a debate between those philosophers who claim that the fundamental fabric of the world is physical, on the one hand, and, on the other, proponents of epistemic arguments aiming to show that there is something essential about conscious states left in obscurity when the world is exclusively considered from the physicalist point of view. We are not going into the details of the arguments and counter arguments in this area. Structural descriptions are enough for our purposes here.

It is a remarkable fact about common epistemic arguments that the job of linking conceivability, understood as a mental act, and metaphysical possibility is usually done by an operation on supposedly self-validating contents of thought. The recognition of validity in this case should be immediately given by the judgement itself. Descartes, Levine and Jackson ground this passage in an intuition of epistemic distinctness that is then supposed to reflect a "real" ontological distinction. It should by now be obvious that the debate can only benefit from the development of formal and

conceptual tools in analytic philosophy. After all, modal logic deepens and clarifies the concepts of metaphysical and epistemic possibilities, which play a central role in epistemic arguments. David Chalmers takes this into account when he develops his novel version of the epistemic argument against physicalism, *the two-dimensional argument against materialism* (aka *the zombie argument*). As he puts it,

> [...] one can legitimately infer ontological conclusions from epistemic premises if one is careful about how one reasons. To do so, the best way is to reason first from epistemic premises to modal conclusions (about necessity and possibility) and from there to ontological conclusions. Here, the crucial issue is the link between the epistemic and modal domains. (Chalmers 2009, p.313)

Chalmers offers a preliminary formulation of the argument:

> [...] it is conceivable that there be a system that is physically identical to a conscious being, but that lacks at least some of that being's conscious states. Such a system might be a zombie: a system that is physically identical to a conscious being but that lacks consciousness entirely. [...] many hold that [zombies] are at least conceivable: we can coherently imagine zombies, and there is no contradiction in the idea that reveals itself even on reflection. [...] From the conceivability of zombies, proponents of the argument infer their metaphysical possibility. [...] From here, it is inferred that consciousness must be nonphysical. If there is a metaphysically possible universe that is physically identical to ours but that lacks consciousness, then consciousness must be a further, nonphysical component of our universe. (Chalmers 2003, p.5f)

Let P stand for all microphysical truths about an actual conscious being (including fundamental physical laws) and Q stand for all the phenomenal truths about this conscious being. Assuming now that physical truths strongly supervene on microphysical truths plus fundamental physical laws, $P \,\&\, {\sim}Q$ describes a being that physically identical to a conscious being, without some of the phenomenal states held by conscious beings. The zombie as conceived by Chalmers satisfies ${\sim}Q$; let's call it a *total zombie*. But the argument can be run with the conception of a *partial zombie*, that is, a being physically indistinct from a normal conscious being, but that doesn't possess at least one of the normal phenomenal states. If we still want to call the argument a "zombie argument", we should also take zombies to include partial zombies.

The distinctive claim of physicalism about the phenomenal is the thesis that phenomenal truths are *necessitated*, i.e., metaphysically entailed, by physical truths, which can be put in formal terms:

$$\Box(P \supset Q),$$

which is equivalent to:

$\sim\Diamond(P \,\&\, \sim Q)$.

Thus, according to the usual semantics of the modal operators, physicalism about the phenomenal amounts to the thesis that any world in which $P \,\&\, \sim Q$ is true is metaphysically impossible. This allows us to render the preliminary formulation of Chalmers' argument in more or less formal terms:

1. $P \,\&\, \sim Q$ is conceivable.
2. $P \,\&\, \sim Q$ is conceivable $\supset \Diamond(P \,\&\, \sim Q)$
3. Physicalism is false.

Later it will be clear why this "argument form" will be developed into a two-dimensional argument. There are many physicalist responses to the argument. One cluster of responses (given by the so-called type-A materialists as Daniel Dennett (1991), David Lewis (1990) and Keith Frankish (2016)) consists in the plain rejection of the conceivability of zombies. For instance, if analytic functionalism[1] is true and physical terms are built into the very meaning of Q, then no system satisfying the description $P \,\&\, \sim Q$ can be a zombie in the sense intended by Chalmers, as it is not *physically identical* to an actual conscious being. If only for the sake of the argument to be developed, we will not take issue with the claim made in premise (1).

Other physicalists react to the two-dimensional argument by accepting the conceivability of zombies, while simultaneously rejecting the link between conceivability and metaphysical possibility. These physicalists (dubbed type-B materialists) usually mobilize the existence of a posteriori necessities as a counterexample to the connection between conceivability and possibility.

We agree with type-B materialism that the key assumption in the two-dimensional argument, as exemplified in premise 2, is that conceivability entails possibility, although we intend to proceed along lines different from what can be normally found in the responses provided by type-B materialists as David Papineau (2002), Brian Loar (1997), and Katalin Balog (2012). We aim to show that Chalmers' very ingenious project of connecting conceivability and possibility with the tools of semantic theory and modal logic fails, as it depends ultimately on seriously circular considerations. We will (i) carefully reconstruct the two-dimensional argument, (ii) show what is at stake in it, and (iii) isolate and take issue with a troubling assumption embodied in any Bridging Principle that needs to be applied to conceivable scenarios in order to generate the metaphysical worlds where zombies could exist. The conclusion shall be that Chalmers' two-dimensional argument against physicalism is at best inconclusive.

2. Scenarios and worlds

The operative idea behind the two-dimensional argument is to infer $\Diamond(P \,\&\, \sim Q)$ (the falsity of physicalism) from the fact that we can conceive $P \,\&\, \sim Q$. This is (part of) what is meant by the slogan "conceivability is a guide to possibility."

Conceivability is no *immediate* guide to possibility. If it were, assuming again Kripke's thesis of the necessity of identities, we could infer the existence of a possible world where water is not H_2O from the fact that we can conceive the sentence *water* $\neq H_2O$. (We are assuming that in whatever reasonable sense in which $P \,\&\, \sim Q$ is conceivable, *water* $\neq H_2O$ is also conceivable.) As previously mentioned, typical cases of rejection that conceivability entails possibility are couched in terms of a posteriori necessities. They aim to show that conceivability is not a safe guide to possibility. Chalmers claims, on the other hand, that a safe passage from conceivability to possibility can be secured by the employment of the two-dimensional framework. As we will see in due time, this is done by an interplay between different dimensions of intensions and possibilities.

Before entering into these details, we need to introduce the notion of scenarios. Chalmers (2011, p. 60) defines a conceivable sentence s as a sentence stating an epistemic possibility, that is, a way things might be for all we know. Alternatively, s is conceivable when a priori reflection cannot show that s is false: when s cannot be ruled out a priori.[2] We can grant that to conceive s is to envision a *complete* scenario in which s is true. Thus, we can accept the thesis Chalmers calls *Plenitude*:

> **Plenitude**: For all sentences s, s is epistemically possible iff there exists a scenario w such that w verifies s (Chalmers 2011, p.64).

Here "scenario" is a technical notion. It must be understood in the sense of a *complete specification* of an imagined situation in which s is true. A scenario is a sort of "world": a maximally specific way things might be for all we know a priori. Thus, to conceive $P \,\&\, \sim Q$ amounts to holding that a zombie "exists" in an imagined "world" very different from our world.

Talk of scenarios (and "worlds" in this epistemic sense) has of course no metaphysical bite. As Chalmers (2002, p. 152) explicitly recognizes, epistemic possibilities "can be regarded as mere intentional objects, useful in characterizing the cognitive or phenomenological structure of modal imagination." It follows from this that we are not "ontologically committed" as long as we reason only with conceivable items and scenarios. Accordingly, imagined scenarios must be distinguished from metaphysically possible worlds. The latter are counterfactual possibilities, i.e., maximally specific ways things might have been. (For the sake of terminological clarity, we will henceforth generically refer to scenarios with the letter v, reserving w, as usual, for metaphysically possible worlds.)

Without the distinction scenario/world, we could get the wrong result that water is not necessarily H_2O. For assuming again that *water* $\neq H_2O$ is not less conceivable than $P \,\&\, \sim Q$, there is a scenario (an intentional object) where water and H_2O are different things. But it would be wrong to infer from this that we are having to do here with a *real* possibility: that water could objectively have gone unaccompanied by H_2O. Indeed, the two-dimensional argument carefully distinguishes the space of scenarios (the space of points where conceivable sentences can be evaluated for truth) from the space of metaphysical, counterfactual possibilities, where sentences can also be evaluated for truth. Truth evaluation in the case of scenarios is what it is called *verification*. *Satisfaction*, on the other hand, is truth evaluation with respect to metaphysical possibilities. Both verification and satisfaction will be given a more or less formal definition below. So far, we have only characterized a notion of sentential truth tied to scenarios. However, as we want to be able to decide whether a sentence *s* is true or false when evaluated in circumstances of evaluation which consist of metaphysically possible worlds, we must establish a link between epistemic possibilities, on the one hand, and metaphysical possibilities, on the other. Put in terms frequently used by Chalmers, we must "construct scenarios in terms of possible worlds."

To yield the intended metaphysical conclusion (that zombies are really possible), the two-dimensional framework puts into operation a principle which makes the epistemic domain correspond to the metaphysical domain. This is what Chalmers has in mind when he talks about "constructing scenarios in terms of possible worlds." Notoriously, Chalmers proposes two ways of constructing scenarios, an epistemic construction and a metaphysical one. The functions associated by Chalmers with both constructions serve different explanatory purposes. The first function should account for the cognitive dimension of meaning, while preserving the realistic, external dimension of mental content. This cognitive aspect of two-dimensionalism does not interest us here. We are concerned only with the prospects of the metaphysical construction of scenarios. This is where the bridging principle comes in.

3. The bridging principle and the metaphysical construction of scenarios

To make good the idea that conceivability can provide a safe guide to real, objective possibility, the two-dimensional argument assumes the availability of a bridging principle (BP), linking the space of intentional/epistemic possibilities constitutively associated with our imaginative capacities, on the one hand, and the admittedly distinct metaphysical space of real possibilities, on the other.

The two-dimensional argument invokes this bridging principle to infer the objective existence of a world where the physical does not determine the phenomenal

from the intentional existence of a "world" where $P \,\&\, \sim Q$ is true. The required bridging principle must be such that the existence of a world where water is not H_2O *does not* follow from an imagined scenario in which *water* $\neq H_2O$.

After going into the details of the bridging principle and the role it plays in the two-dimensional argument, we will take issue with the grounds offered as its support.

Let $ver(v,s)$ be a relation between an imagined scenario v and a sentence s, which obtains when v *verifies* s, i.e., when ideal, a priori reflection on v reveals it as a situation in which s is true.

Let $sat(w,s)$ be a relation between a metaphysically possible world w and a sentence s, which obtains when w *satisfies* s, i.e., when the following subjunctive conditional is a priori:

> If w had obtained, it would have been that s.

The verification and the satisfaction of a sentence s correspond to different forms of semantic evaluation. First, the circumstances of evaluation in each case are categorically distinct: imagined scenarios in one case, metaphysical worlds in the other case. Second, the evaluation associated with satisfaction, but not the one associated with verification, involves a subjunctive conditional.

Although scenarios *qua* intentional objects are conceptually distinct from counterfactually possible words, Chalmers claims that any such world *w considered as actual* can play the role of a maximally specific epistemic possibility. This means that we can take any possible world w as a hypothesis about what our world is like. As Stephen Yablo (2002, p.449) remarks, "we do not in general *believe* the hypothesis." But we can always evaluate a sentence s with respect to a world w considered as actual by asking whether s holds on the hypothesis that our world turns out to be w. The crucial point is that we cannot rule out a priori that our actual world is w. For instance, it is only a posteriori that we can rule out that the world we are living in turns out to be Putnam's *Twin Earth*, a possible world where the "watery stuff" is not H_2O. Thus, *Twin Earth* considered as actual works as a maximally specific epistemic possibility, a scenario which is, for instance, in the verification relation to the conceivable sentence *water* $\neq H_2O$. Scenarios resulting from the consideration of metaphysically possible worlds as actual cannot be regarded as mere intentional objects, as they are somehow "constructed" with the material of real, objective possibilities, which are *extensionally* individuated possibilities. We will henceforth use $[w]$ to denote a world *w considered as actual*.

We can now state more precisely the bridging principle:

> **The Bridging Principle (BP):** Imagined scenarios determine metaphysically possible worlds, so that to each imagined scenario v verifying a conceivable s there is a possible world w_v, which, once considered as actual, works as a maximally specific epistemic situation where s holds.

As per BP, it must be the case that when we conceive P & $\sim Q$, we are ultimately envisioning a *real* possibility. We then take this possibility as actual, we put ourselves into the position to epistemically evaluate P & $\sim Q$ for truth—in the same sense in which we ultimately envision *Twin Earth* when we conceive *water* $\neq$ H_2O. And in *both* cases, the evaluation renders the same value 1. Against the background of Twin-Earth *considered as actual*, water is different from H_2O. The divergence in the respective truth evaluations of P & $\sim Q$ and *water* $\neq$ H_2O (hopefully!) emerges when it comes to testing the subjunctive satisfaction of these sentences by the corresponding worlds.

We can now formulate the verification relation by replacing v, which designates a scenario, with $[w]$, the expression for a world w considered as actual. Thus, "$ver([w],s)$" designates a relation which obtains when the world w considered as actual verifies a conceivable sentence s. The relation $ver([w],s)$ obtains just in case the indicative conditional "When w is actual, then s" is true. Arguably this indicative conditional should be interpreted as a material implication (cf. Yablo 2002, p.450f, n6). Thus, $ver([w],s)$ is true just in case the following material conditional is true:

If d, then s,

where d is a complete description of w. We will come later to the constraints Chalmers imposes on d.

4. Putting some pieces together

Let the conceivable sentence s be *water* $\neq H_2O$. By referring to *Plenitude*, we can say that there is a scenario v that verifies the hypothesis that water is a different stuff from H_2O:

$ver(v, water \neq H_2O)$ is true.

Assuming the validity of BP, corresponding to this scenario v there is a metaphysically possible world w_v which, when considered as actual, verifies the idea that water and H_2O are different things:

$ver([w_v], water \neq H_2O)$ is true.

Intuitively, a good candidate for w_v in this case is *Twin Earth*: a possible world where the watery stuff is *XYZ* ($\neq H_2O$). But *Twin Earth* is a counterfactually possible world where water still *is* H_2O: if *Twin Earth* had obtained, it would not have been the case that *water* $\neq$ H_2O. Thus, we have:

$sat(TwinEarth, water \neq H_2O)$ is false.

The conjunction of $ver([TwinEarth], water \neq H_2O)$ and the negation of $sat(Twin\ Earth, water \neq H_2O)$ reflects what we independently expect: *water* $\neq$ H_2O may be conceivable, but it does not represent a real metaphysical possibility.

Compare this with the case where the conceivable s is $P \,\&\, {\sim}Q$. *Plenitude* allows us to state

$ver(v, P \,\&\, Q)$ is true.

Assuming again the validity of BP, there is metaphysically possible world w_v which, when considered as actual, verifies $P\&{\sim}Q$. So we have:

$ver([w_v], P \,\&\, Q)$ is true.

Now a distinctive aspect of $P \,\&\, {\sim}Q$, which is *not* shared by *water* $\neq$ H_2O, ensures, as we will presently see, that the very same world w_v that verifies $P \wedge {\sim}Q$, when it is considered as actual, also satisfies it:

$sat(w_v, P \,\&\, Q)$ is true.

The conjunction of $ver([w_v], P \,\&\, Q)$ and $sat(w_v, P \,\&\, Q)$ means that zombies are not merely conceivable; they are really possible.

5. Varieties of intensions

To characterize the distinctive aspect of $P \,\&\, {\sim}Q$ responsible for the "right" result (the objective possibility of zombies, but not of *water* $\neq$ H_2O), we must first distinguish the intensions that can be associated with any sentence s. The two-dimensional framework allows the distinction between three kinds of intensions:

- The (purely) epistemic intension of s maps imagined scenarios v to the truth-value of $ver(v, s)$.
- The primary intension of s is a mapping from worlds w to the truth-value of $ver([w], s)$. It is grounded in the epistemic, indicative evaluation of s in worlds considered as actual.
- The secondary intension of s is a function from worlds w to the truth-value of $sat(w, s)$. It returns the result of the counterfactual, subjunctive evaluation of s in worlds w.

Analogous definitions can be given for the case of sub-sentential terms.

If we adopt the two-dimensional representation usual in double-index semantics and replace contexts of utterance with worlds w considered as actual, while keeping

worlds *w* considered as counterfactual as circumstances of evaluation, then the result is that the primary intension is diagonal, while the secondary intension is horizontal. Consider, for instance, the following simplified two-dimensional semantic representations of *water* (the leftmost matrix) and of *water* $\neq H_2O$ (the rightmost matrix), where as usual the actual world is represented by @, while *i* represents Twin-Earth.

$$\begin{bmatrix} & @ & i \\ @ & H_2O & H_2O \\ i & XYZ & XYZ \end{bmatrix} \qquad \begin{bmatrix} & @ & i \\ @ & 0 & 0 \\ i & 1 & 1 \end{bmatrix}$$

It is accepted by virtually everyone that the non-logical terms composing *Q*, which are exclusively phenomenal terms, ensure that its primary and secondary intensions coincide.[3] This reflects an intuition held by classical conceptions of the items in our mind, according to which they are as they appear to us. Also, Kripke claims that it is necessarily pain what I have when it seems to me that I am having pain, while at least sometimes it is not true that what is epistemically indistinguishable from water (what appears to be water) is water. Putting in terms of the 2D semantics, the epistemic evaluation of phenomenal terms does not change when we change the world considered as actual. It follows from this that the distribution of 1's and 0's along the diagonal of the matrix representing the semantics of *Q* perfectly mirrors the corresponding distribution along the horizontal: the primary intension of *Q* coincides with its secondary intension. Moreover, many philosophers hold that *mutatis mutandis* this also applies to the non-logical terms which appear in the composition of *P*. Fundamentally physical terms, it is claimed by many, have constant extensions across different worlds considered as actual. Thus, both *P* and *Q*, as well as any combination thereof with the devices of propositional logic, would have coinciding primary and secondary intensions. As we will presently see, Chalmers himself disagrees with the coincidence thesis as applied to *P*. As regarding *Q* Chalmers is definitely in agreement with those philosophers who claim that its primary and secondary intensions are the same.

The opposite view regarding the behavior of *P* in different worlds considered as actual is held by other philosophers, Chalmers himself among them. This is a plausible view if we accept (i) that the primary intension of any physical term is tied to the role it plays in a theory, and (ii) that the theory in question describes the structure of the physical world, being silent on the non-relational nature of the items so structured. The secondary intension of a physical term, in its turn, is tied to the properties that actually play the role identified by the theory. Thus, the primary intension of H (the term for hydrogen), for instance, points in the first place to the hydrogen role as identified in the best theory of physical reality, while the secondary intension of H points to what plays the hydrogen role *in the actual world*. Plausibly, H is a *rigid*

designator: its extension as fixed in the actual world is constant across counterfactual possibilities. But this does not prevent that the counterfactually constant extension varies with scenarios, i.e., with worlds considered as actual. In the terminology of two-dimensional semantics, physical terms are rigid designators, but, unlike phenomenal terms (interpreted along the lines of the last paragraph), they are not *superrigid designators*. Of course, all this follows from the interpretation of physics as an account of the structure of reality, not of its intrinsic nature. Philosophers who accept this interpretation, regardless of what they think about the functioning of the terms comprising Q, must deny that the primary and the secondary intensions of $P \,\&\, {\sim}Q$ coincide.

As it will be presently clear, the two-dimensional argument against physicalism does not go through without the semantic thesis that $P \,\&\, {\sim}Q$ behave the same regarding both the primary and the secondary intensions. That is why we will not consider here the alternative view implied by the structuralist interpretation of physics. The crucial point for our purposes is that only under the supposition that the primary and the secondary intensions of $P \,\&\, {\sim}Q$ coincide, the satisfaction of $P \,\&\, {\sim}Q$ by a certain world w is guaranteed by its being in the verification relation to this very same world w (provided, of course that it is considered as actual), which should lead to the intended result: physicalism is false. This leaves intact the result obtained in the case of water and H_2O: the primary intension of *water* $\neq H_2O$ is different from its secondary intension, and verification does not entail satisfaction.

6. A critical appraisal

The following makes explicit the steps of the two-dimensional argument:

1. $P \,\&\, {\sim}Q$ is conceivable.
2. As per *Plenitude*, there is a scenario v such that $ver(v, P \,\&\, Q)$ is true.
3. As per BP, the scenario v determines a world w_v such that $ver([w_v], P \,\&\, Q)$ is true.
4. The primary and the secondary intensions of $P \,\&\, {\sim}Q$ coincide.

Therefore,

5. $ver([w], P \,\&\, Q)$ entails $sat(w, P \,\&\, Q)$.

Therefore,

6. $sat(w_v, P \,\&\, Q)$ is true.

Therefore,

7. $\Diamond(P \,\&\, \sim Q)$

Clearly, the key element here is BP, which "turns" an epistemically conceived scenario v into a possible world w_v. The latter (provided it is considered as actual) functions as a context for the epistemic evaluation of $P \,\&\, \sim Q$ and should prove farther down the line to satisfy $P \,\&\, \sim Q$. As to what it is exactly for a world (considered as actual) to verify a sentence, Chalmers (2011, p.68) writes: "We could then say that a world w verifies a sentence token s when d implies s, where d is a canonical specification of w." The canonical specification of a possible world, in its turn, is an infinitary sentence in a highly idealized language. It is framed exclusively in *neutral* terms (besides, of course, logical terms), i.e., expressions that do not change their extension when used by my twin in a Twin-Earth situation. Neutral terms in this sense are called by Chalmers "Twin-Earthable". Intuitively, "water" is Twin-Earthable, while "H_2O" and "*XYZ*" are not. According to Chalmers, not only "computer," "philosopher," and fundamental-physical terms, but also "consciousness," and phenomenal terms like "pain" are neutral. In a more traditional vein, neutral terms could be said to designate entities whose mode of appearance is undissociated from their essence. To avoid the metaphorical tone of this explanation, Chalmers offers a more technical definition, according to which neutral expressions behave "the same with respect to both verification and satisfaction" (Chalmers 2011: 71), which means that they have coinciding primary and secondary intensions, or alternatively that they are super rigid. This allows for an alternative formulation of the verification relation: "w verifies s if a canonical specification of w epistemically necessitates s" (Chalmers 2011, p.70).

It is also as given by its canonical specification that the world w_v will be tested for the counterfactual satisfaction of $P \,\&\, \sim Q$, the test consisting in determining whether $P \,\&\, \sim Q$ *would* be true if the corresponding canonical specification d of w_v *were* true. Generalizing, Chalmers (2011, p.70) writes: "w satisfies s if a canonical specification of w metaphysically necessitates s." Although at this point we cannot yet put w_v to the test mentioned above, it is important to keep in mind that the languages of d and of $P \,\&\, \sim Q$ are not distinct: they contain (besides logical terms) only neutral terms.

It now seems that, under the assumptions made by the two-dimensional argument, getting the intended word w_v is cheap. We start from the apparently unproblematic assumption that to conceive $P \,\&\, \sim Q$ basically means to envision a maximally specific epistemic possibility, a scenario v in which $P \,\&\, \sim Q$ is true. The specification of v can only be an infinitary sentence. Moreover, it must be a non-contradictory sentence. Next, we translate, if only in principle, this sentence into one that contains only neutral expressions plus logical terms. The neutral terms here are purely physical: after all, we are specifying a scenario in which *ex hypothesi* zombies "exist." This sentence is not true, as there are no zombies in the actual world—or so we can assume. But it is very natural to wonder whether $P \,\&\, \sim Q$ would be true if this sentence were true.

Notice that simply thinking about what would follow from a scenario-specifying sentence if it were true amounts to surreptitiously treating this sentence as the canonical specification of a metaphysical world. That is, we are implicitly *presupposing* that "out there" in the metaphysical space there is a possible world waiting to be so specified. And thinking that P & $\sim Q$ is satisfied by this presupposed world means treating it as a zombie world, the complete metaphysical possibility which verifies the zombie sentence. If we are not careful here, we may fall prey to the illusory impression that simply by conceiving a zombie we reach a metaphysically possible zombie world. But the fact is that no purely physical world is *ipso facto* a zombie world. Being purely physical may be a necessary condition for being a zombie world in that sense, but it is not sufficient. Most importantly, the existence of the relevant counterfactual possibility has been *presupposed* in the above reasoning from epistemic considerations. We cannot make a metaphysical possibility out of the thin air of scenarios. We can only say that the zombie we are conceiving could well be a genuine denizen of the possible world canonically specified by the same sentence specifying the zombie scenario, *provided that such a world is really possible*.

Let us then ask: what could entitle us to the assumption that there is in the space of metaphysical possibilities a world that can be canonically specified by a sentence originally stating an epistemic possibility? At this point in the dialectic, Chalmers introduces

> ***Metaphysical Plenitude*:** the thesis that for all sentences s, s is epistemically possible iff there exists a centered world that verifies s. (Chalmers 2011, p.71)

Similarly, Levine (2018) claims that the "limits of possibility" are defined by the a priori sources of logic and the structure of our concepts. It follows from this that if a sentence contains neither conceptual relations nor logical relations that could impose limiting constraints on the metaphysical space, that is, if a sentence is conceptually and logically coherent, then a counterfactual possibility corresponds to this sentence. To accept this is to endorse the idea that there is nothing "arbitrary, or gappy about possible world space," which is how Levine interprets "plenitude" (Levine 2018, p.54). In particular, he agrees with Chalmers that, given the absence of conceptual or logical connections between descriptions of physical facts and descriptions of phenomenal facts, the conceivability of a zombie points to a real possibility (Levine 2018, p.55).

Compare with this what some philosophers of science assert about the metaphysical necessity of natural laws. For instance, referring to the law according to which the intensity of light from a constant source falling on an area is inversely proportional to d^2, where d is the distance between the source and the area, Alexander Bird (2005) argues that it would be wrong to think that there might be a possible world

in which the light intensity is proportional to $d^{-2,000001}$. The crucial point made by Bird is this:

> [...] a world in which the intensity is proportional to $d^{-2,000001}$ is not at all similar to ours; it is one where energy (or mass-energy) is not conserved (and it is not clear to me that such a world is genuinely possible). (Bird 2005, p.365)

Put in the terms mobilized by Chalmers and Levine, Bird is (cautiously) claiming here that metaphysical plenitude is false, that there is a gap in possible world space. As the *experimentally discovered*, relevant constant here (the exponent of displacement) is 2, a scenario that verifies the hypothesis that it is different from 2 is OK. But we cannot be assured, Bird says, that corresponding to this epistemic possibility there is a genuine metaphysical possibility. In other words: there is arguably no world that, once considered as actual, could play the role of the scenario *vis-à-vis* the conceivability of a situation in which the exponent of displacement differs from 2.

Cases like this are called by Chalmers *strong a posteriority necessities* (or simply *strong necessities*). As opposed to standard a posteriori necessities like *water* $= H_2O$, a *strong necessity* is characterized by the fact that it is verified by all genuinely possible worlds. According to Chalmers, a list of these at least potentially problematic cases for *Metaphysical Plenitude* includes inter alia (i) the view (exemplified above) that fundamental laws are metaphysically necessary, (ii) the materialist view, according to which phenomenal truths are necessitated by, but not a priori derivable from, physical truths, and (iii) the view that certain mathematical claims are true and necessary, but are not knowable a priori (cf. Chalmers 2011, p.72).

Now, *Metaphysical Plenitude* is a bi-conditional. The truth of the right-to-left conditional can be taken for granted. We can infer the epistemic possibility of s from the existence of a possible world which verifies s (provided it is considered as actual). The left-to-right conditional, on the other hand, has the same content as BP, whose validity is in question here. How can we guarantee that there really is in the metaphysical space a world w_v such that its canonical specification matches up the proper specification of a scenario v which verifies s? *Metaphysical Plenitude* by itself does not solve our problem. At best, it merely reformulates it. What the two-dimensional argument needs here is an extra argument supporting *Metaphysical Plenitude* and, by extension, BP. The supporting argument could be indirect, i.e., it could proceed by the refutation of all those views implying strong necessities. Indeed, Chalmers aims "to deny that there are any strong necessities" (Chalmers 2011, p.72). However, at the end of the day he prefers to "argue in reverse." He writes: "the fact that the link between epistemic possibility and verification by possible worlds is so strong elsewhere gives reason to believe that these claims [to the existence of strong necessities] are incorrect" (Chalmers 2011, p.72).

It is hard to avoid the feeling that at this point the two-dimensional argument runs in circles. The argument invokes the core of the *Metaphysical Plenitude* thesis, namely BP, to justify the claim that there is in metaphysical space a world that verifies the epistemic possibility of a zombie. You cannot treat the simply asserted absence of a gap at *this* spot of metaphysical space as good evidence that there are no gaps in other spots, marked by the purported counterexamples of strong necessities. This would amount to grounding *Metaphysical Plenitude* in itself.

A possible way out requires that we take a leap of faith, thereby weakening the two-dimensional argument. Let us ask how far we can go if we simply assume without further justification that there is a world w_v such that $ver([w_v], P \,\&\, Q)$ is true. We are immediately entitled to add a row and a column to the matrix representing the semantics of $P \,\&\, \sim Q$ and to enter 1 into the corresponding cell in the diagonal. Next, we take into account the already established fact that the secondary intension of $P \,\&\, \sim Q$ exactly mirrors its primary intension. This leads to a 1 where that @-row and the w_v-column intersect, which is the result we expected: the world w_v, whose existence we assumed, not only verifies $P \,\&\, \sim Q$, but also satisfies it. But this will not do as an explanation. We need metaphysical worlds in the first place to get primary and secondary intensions (and the associated notions of verification and satisfaction). If the worlds in question must be given to us in their canonical specifications restricted to neutral expressions, and if these expressions are characterized in terms of the interplay between their primary and secondary intensions, then we are moving again in a circle. This is of course no direct argument against the existence of the world w_v assumed by the two-dimensional argument. We are only pointing to the not irrelevant fact that the assumption results in a likely vicious circularity.

7. Conclusion

Our goal here was to make troubles about BP explicit. First, Chalmers' attempt to establish a relation between an imagined scenario and a non-intentionally individuated, objective world, so that both the scenario and the world epistemically verify a certain conceivable s begins with the idea of a world w whose canonical specification d implies s. This seems to *presuppose* that there is a possible world corresponding to every circumstance that can be consistently conceived. Second, the two-dimensional argument embodies the idea that when we are conceiving any (non-contradictory) sentence s, we are envisioning an object of thought, a scenario v, but also somehow, and more fundamentally, we are devising a metaphysically possible world w_v which (i) verifies s, when considered as actual, and (ii) can be tested for the counterfactual, subjunctive satisfaction of s. As we have seen, it is the job of BP to deliver this world. In this case, however, we cannot begin the explanation of BP with the asser-

tion that *there is* a real world corresponding to any conceivable scenario. That would be evidently circular.

If the remarks above are sound, there are no good reasons for the *generalized* assumption of the Bridging Principle. We cannot be sure that there is a way to go beyond the boundaries of imagination in all cases that may interest us—no general way to get to the required objective world starting from the intentional objects of our imaginative conceptions, even under the assumption that these conceptions are logically and conceptually consistent and proceed under idealized conditions. We cannot rule out the existence of counterexamples based on strong necessities. In the particular case of conceiving that *water* $\neq H_2O$, it is plausible that the role played by the imagined scenario in the epistemic verification of water's not being H_2O can be played by an unsuspected metaphysically possible world, which we can independently describe. It is a world where the "watery stuff" is, for instance, *XYZ*, a stuff different from H_2O. Accordingly, there is in this case a robust distinction between the primary intension and the (purely) epistemic intension of the conceivable sentence. In the case of $P \,\&\, {\sim}Q$, however, we have no independent description of the objective world that should deliver the epistemic context for the verification of the zombie hypothesis. We seem to be forced to say that the posited world w_v is simply the zombie world and that it simply corresponds to the imagined zombie scenario v with which we started the reasoning along the lines of the two-dimensional argument. It is very difficult to avoid the thought that the primary intension in this case is just the (purely) epistemic intension "in sheep's clothing." Replacing "v" with "$[w_v]$" in the expression of the verification relation is no guarantee that "w_v" refers to a real world.

Looked at this way, the two-dimensional argument comes close to assuming as a premise what should be its conclusion. It does not show, as it should, that the objective order (the space of metaphysical possibilities) contains a world corresponding in the right way to the imagined scenario verifying $P\&{\sim}Q$. In a relevant sense, the two-dimensional argument against physicalism more assumes then proves that a zombie is a real possibility. But this means that it is at best inconclusive.

References

Balog, K. 2012. In Defense of the Phenomenal Concept Strategy. *Philosophy and Phenomenological Research* 84(1): 1–23.

Bird, A. 2005. The Dispositionalist Conception of Laws. *Foundations of Science* **10**(4): 353–70.

Chalmers, D. 2002. Does Conceivability entail Possibility? In: T. Gendler & J. Hawthorne (ed.), *Conceivability and Possibility*, p.145–200. Oxford: Oxford University Press.

Chalmers, D. 2003. Consciousness and its Place in Nature. In: S. Stich & T. Warfield (ed.), *Blackwell Guide to the Philosophy of Mind*, p.102–142. Oxford: Blackwell.

Chalmers, D. 2009. The Two-Dimensional Argument Against Materialism. In: B. McLaughlin & S. Walter (ed.), *Oxford Handbook to the Philosophy of Mind*, p.313–338. Oxford: Oxford University Press.

Chalmers, D. 2011. The Nature of Epistemic Space. In: A. Egan & B. Weatherson (ed.), *Epistemic Modality*, p.60–107. Oxford: Oxford University Press.

Dennett, D. 1991. *Consciousness Explained*. New York: Back Bay Books.

Frankish, K. 2016. Illusionism as a Theory of Consciousness. *Journal of Consciousness Studies* **23**(11-12): 11–39.

Howell, R. 2013. *Consciousness and the Limits of Objectivity: The Case for Subjective Physicalism*. Oxford: Oxford University Press.

Jackson, F. 1982. Epiphenomenal Qualia. *Philosophical Quarterly* 32(April): 127–136.

Kripke, S. 1971. Identity and Necessity. In: M. K. Munitz (ed.), *Identity and Individuation*, p.135-164. New York: New York University Press.

Levine, J. 1983. Materialism and Qualia: The explanatory gap. *Pacific Philosophical Quarterly* **64**(October): 354–61.

Levine, J. 2018. Bruteness and Supervenience: Mind vs. Morality. In: C. Mekios & E. Vintiadis (ed.), *Brute Facts*, p.45–62. Oxford: Oxford University Press.

Lewis, D. 1966. An Argument for the Identity Theory. *Journal of Philosophy* 63: 17-25.

Lewis, D. 1990. What experience teaches. In: W. Lycan (ed.), *Mind and Cognition*, p.29–57. Oxford: Blackwell.

Loar, B. 1997. Phenomenal States (Revised Version). In: O. Flanagan; N. Block; G. Guzeldere (ed.), *The Nature of Consciousness*, p.597-616. Cambridge, Massachusetts: MIT Press.

Papineau, D. 2002. *Thinking About Consciousness*. Oxford: Oxford University Press.

Schroer, R. 2013. Do the Primary and Secondary Intensions of Phenomenal Concepts Coincide in all Worlds? *Dialectica* **67**(4): 561–577.

Yablo, S. 2002. Coulda, Woulda, Shoulda. In: T. Gendler & J. Hawthorne (ed.), *Conceivability and Possibility*, p.441-492. Oxford: Oxford University Press.

Notes

[1]As conceived by Lewis (1966), analytic functionalism says that psychological states in general, and conscious states in particular, are individuated (constituted) by their causal-functional, ultimately physical profile.

[2] This is Chalmers' notion of *negative* conceivability. He proposes also a *positive* notion. The argument can be discussed independently of the chosen notion. We prefer the negative one.

[3] There seems to be in the relevant literature only one paper that takes issue with the thesis that the primary and secondary intensions of *Q* coincide, namely, "Do the Primary and Secondary Intensions of Phenomenal Concepts Coincide in all Worlds?" by Schroer (2013).