Mirza Mehmedović

Possible Worlds, Zombies and Truth Machines

1. *What does it mean to have a zombie twin?*

In philosophical terms a twin zombie (TZ) is an individual identical to a conscious human being, in this case it is my/your identical twin, but without conscious experience such as the perception of color and scent. It is not an individual that actually exists in the universe, but rather a possible individual, i.e. a subject which has all the other characteristics of the actual universe. In other words, a TZ is a possible individual micro-physically identical to me and at the same time devoid of conscious experience. As Iris Murdoch wrote[1] in a discussion on behaviorism, "all is silent and dark within." It should also be added that while a TZ differs from me only due to its lack of conscious experience, its observable behavior is nonetheless identical to mine; a TZ answers questions and acts in an apparently rational manner in the surrounding environment. In the end, if we admit that a zombie is a coherently conceived "object," by extending this reasoning further, it is possible to imagine worlds inhabited by zombies where there is no consciousness.

One of the premises of the zombie argument is that the conceivability of zombies implies their metaphysical possibility[2]. The topic of the conceivability of zombies comes as a logical consequence of the idea developed by David Chalmers[3] in "The Conscious Mind" indicating that there are certain features of reality which are not reducible to physical properties. If proven true, this argument (or supervenience argument) falsifies physicalism[4], the doctrine which holds that all properties of a superior level in respect to the base properties

[1] I. Murdoch, *The Sovereignty of Good*, Routledge and Kegan Paul, London 1970, p. 13.

[2] D. J. Chalmers, *Does conceivability Entail Possibility?*, in T. Gendler - J. Hawthorne (ed.), *Conceivability and Possibility*, Oxford University Press, New York 2002, pp.145-200.

[3] Id., *The Conscious Mind: In Search of a Fundamental Theory*, Oxford University Press, New York 1996.

[4] Physicalism is a philosophical doctrine introduced, among others, by the Austrian philosopher Otto Neurath. See: *Otto Neurath and the Unity of Science*, in J. Symons - O. Pombo - J. Manuel Torres (eds.), *Logic, Epistemology, and the Unity of Science 18*, Springer, Dordrecht 2011.

which are described by physics, are ultimately reducible to the latter according to certain principles of reduction or "bridge laws". Conversely, Daniel Stoljar notes[5] that if physicalism is characterized as the doctrine that, "for any subjects S and S* and worlds W and W*, if S in W and S* in W* are physically identical, then they are psychologically identical;"[6] so if this doctrine is true, then it is impossible for me to have a twin zombie.

Chalmers[7], along with a number of authors[8], openly supports the consistency of zombies; while others[9], including Eric Marcus[10] and Robert Kirk[11], have attempted to demonstrate that zombies are logically impossible via various arguments. For example, Stoljar notes[12] that physicalists generally deny that conceivability is a sure guide to the possibility. (Towards the end of this paper, after presenting my argument, I comment on one of these criticisms around the conceivability of zombies, namely that of Robert Kirk, explaining why in my opinion it does not completely hit the mark.)

1.2. *Supervenience and types of supervenience*

David Chalmers featured two types of supervenience: "logical," that is characterized as being metaphysically necessary or necessary in all possible worlds, and "natural" supervenience which we can also call "contingent" supervenience. According to Chalmers, consciousness is a contingently supervenient (i.e. naturally supervenient) feature of real individuals, to the extent that the individuals of this universe are real biological systems with conscious experience; but not logically supervenient since it is conceivable a consistent world,

[5] D. Stoljar, *The Conceivability Argument and Two Conceptions of the Physical*, in «Philosophical Perspectives» 15(2001), pp. 393-413.

[6] *Ibi*, p. 393.

[7] D.J. Chalmers, *The Two-Dimensional Argument against Materialism*, in *The Character of Consciousness*, Oxford University Press, New York and Oxford 2010.

[8] N. Block, *On a Confusion about a Function of Consciousness*, in «Behavioral and Brain Sciences» 18(1995), pp. 227-247; N. Block, *The Harder Problem of Consciousness*, in «Journal of Philosophy» 99(2002), pp 391-425; J. Levine, *Purple Haze: The Puzzle of Consciousness*, Oxford University Press, Oxford-New York 2001.

[9] For a review of the critiques see: http://consc.net/responses.html

[10] E. Marcus, *Why Zombies are Inconceivable*, in «Australasian Journal of Philosophy» 3(2004), pp. 477-490.

[11] R. Kirk, *Why There Couldn't Be Zombies*, in «Proceedings of the Aristotelian Society» Supplementary Volume 73(1999), pp. 1-16; see also: K. Frankish, *The anti-zombie argument*, in «Philosophical Quarterly» 57(2007), pp. 650-666.

[12] D. Stoljar, *The Conceivability Argument and Two Conceptions of the Physical*, cit., p. 394.

physically identical to ours, but in which individuals would have no conscious experience. Let us see what this means.

According to the general definition provided by Chalmers[13], "supervenience" is a relationship between two sets of properties. The higher-level properties (or HLPs) are supervenient over the properties of the basic, or lower level (LLPs). The LLPs that usually are taken into account are those basic to physics which, based on the physicalistic approach, tend to determinate all the HLPs from those of biology or chemistry. A stricter definition states that the HLPs are supervenient over LLPs if there are not two identical situations in respect to the LLPs but that differ with respect to the HLPs. Thus, if we speak of complex organisms, we will say that any two objects with the same physical properties are also biologically identical (i.e. indiscernible[14]). Moreover, since the properties covered by biology do not have a distinct ontological problem, in this case it could be stated that the relationship between LLPs and HLP, if true, is true in all possible worlds.

However, as we said it is possible to distinguish two types of supervenience, which roughly correspond to two kinds of relationships between sets of properties. The logical supervenience applies to those HLPs that are reducible to the LLPs in all possible worlds. In other words, it is not possible to imagine a situation where that type of relationship would not be valid or may differ with respect to some element, for example, it is said to be metaphysically impossible (not even God could do it) for male *hens* to exist. This defines the relationship of logical supervenience in terms of *a posteriori* necessary identities. For example, knowing that "water is H_2O" is true, assuming this as an *a posteriori* identity, we come to the conclusion that worlds in which water is not H_2O (Saul Kripke's argument[15]) are not possible. The same goes for known identities such as "Hesperus is Phosphorus" or "Tullius is Cicero". As for the example of hens, we say that a hen has a kind of internal structure discovered *a posteriori* thereby fixing the reference to the term "hen" in all possible worlds. The same generally applies in nature to all the natural kinds[16]. (We are obviously not interested in

[13] D.J. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory*, cit., pp. 30 ss.

[14] D. Lewis, *Papers in Metaphysics and Epistemology*, Cambridge University Press, New York 1999; On Lewis' materialist thought, see: S. Guttenplan, *A Companion to the Philosophy of Mind*, Blackwell Publishers, Oxford 1994.

[15] S.A. Kripke, *Naming and Necessity*, Blackwell, Oxford 1972.

[16] S.A. Kripke, *Reference and Existence. The John Locke Lectures*, Oxford University Press, New York 2013. Kripke also argues that the names of imaginary beings are not real names, but only "pretended names", since these names do not apply to things having a discernible internal structure (names of mythological animals as "Unicorn" or "Pegasus" have no reference *tout court*).

going into depth about the reference problems usually discussed, as in some counterfactual situations where such names could not refer to the same object.)

Chalmers correctly says that logical supervenience is stronger than natural supervenience and that the latter is characteristic of a subset containing the set of objects that satisfy the occurrence of the stronger type. The more interesting epistemological argument captures the intuition that natural supervenience (that which always and systematically occurs in the form of a close correlation between property sets) leaves room for alternatives not currently realized or realizable.

The clearest example used by Chalmers is the equation of the state of perfect gases, according to which we know that the pressure exerted by one mole of a gas depends on the temperature and occupied volume. Knowing the temperature (T) and occupied volume (V), we can always determine the pressure (p) exerted by the gas; arriving at the equation of pV = nRT, where (R) is a certain constant. Chalmers notes[17] that this law only correlates in respect to the facts of nature currently observable and that with a slightly higher or lower R which is equal to other factors (temperature and volume), the pressure would be different. Similarly, we say that it is logically possible for two cubes to exist; one made of gold and the other of uranium-235, both measuring a kilometer on each side, yet we also know that this is impossible in nature due to the instability of the uranium-235 atoms. Therefore, it is generally stated that logical possibilities include natural ones but not *vice versa*. So although it is possible for a thousand monkeys to type out Hamlet, it is extremely unlikely that they will actually succeed.

The problem that now presents itself is how to determine which category of supervenience is within experience; that is, what is the nature of the phenomenon of "consciousness". Entering into the merits of the problem of mind-brain identity, Chalmers states that the case of consciousness does not seem to fit in cases of logical supervenience; that is, in relation to an *a posteriori* necessary identity between sets of properties, for which we have the basic properties of physical reality in one set and properties of conscious experience in the other (as in the subjective experience of colors or flavors). Rather, Chalmers continues, it seems that consciousness is only naturally supervenient. This is true, but it is not necessarily true that bodies like ours have conscious experience. Logically speaking, there may be worlds in which (with the same physical, biological and psychological properties that do not present problems of ontological reducibility to physical characterization) there would be no conscious experience.

[17] D.J. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory*, cit., p. 33.

1.3. *Supervenience and mind-body identity*

The irreducible problem that philosophy of mind raises, according to Chalmers, is when we attempt to treat consciousness in physicalistic terms[18]. The key issue under discussion is that the characteristics of consciousness, namely qualia, do not seem in any way necessarily supervenient to the characteristics of biological and physical individuals in the real world. In a sense it is trivially true that a real individual in normal conditions has conscious experience. Nevertheless, this does not seem to go beyond the mere finding that our physical organization (namely our cortical tissue) implements the conscious experience. But as suggested, things could have gone differently, so that we might have developed a modular 3D brain-like cortical tissue, constructed with a silk protein-based scaffold, ECM composite and primary cortical neurons. On this point Chalmers states one of his most controversial theses; since there is no strong scientific reason for stating (as an *a posteriori* necessary truth) that the properties of consciousness are logically supervenient to our biological and, ultimately, physical characteristics. However, it is consistent to imagine possible worlds in which individuals identical to us would be completely devoid of conscious experience[19]. The type of conceptual consistency that Chalmers calls into play is analogous to the recognition of a possible world where "water" (that watery stuff, the colorless liquid that fills the oceans) would prove to be "XYZ" instead of "H_2O". It is essentially the ability to conceive non-standard realizations of our functional[20], organizational characteristics. Therefore, if alternative material characterizations are conceivable

[18] It should keep in mind that the term consciousness is used to uniquely identify the conscious experiences with phenomenal properties, such as colors, shapes, sounds and so on. The broader concept "mind" is used to designate the set of cognitive processes (both conscious and unconscious).

[19] *Ibi*, pp. 84-88.

[20] See the classical thought experiment described by: N. Block, *Troubles with functionalism*, in C.W. Savage (ed.), *Perception and Cognition: Issues in the Foundation of Psychology*, University of Minnesota Press, Minneapolis 1978. Reprinted in N. Block, *Readings in the Philosophy of Psychology*, Vol. 1, Harvard University Press, Cambridge (MA) 1980. Block imagines what would happen if the functional organization of the brain was reproduced by the entire Chinese population, where each inhabitant corresponded to a neuron and each radio link to a synapse. This type of experiment, however, is not new. Leibniz in his *Monadologie* imagined what would happen if we had a huge machine structurally capable of thinking, in which we can get in as in a mill. In both cases the underlying idea is that – apart from the gears or the elements constitutive of the organization of the physical system in question – from any particular material condition would (logically) follow that that system would necessarily be conscious. We would not find anything in it that explains subjective experience; See G.W. Leibniz, *Monadologie* (1714).

for certain primary intensions[21], then the relationship between primary intensions and extensions appears to be contingent.

The problem is that while we have an idea of the criterion by which we would test whether the liquid in question is H2O or XYZ, in the case of zombies, that seems more controversial. Is it possible to think of a way in which we could, rather than simply reaffirm that zombies seem logically consistent, strengthen or weaken this conviction on the basis of certain other basic beliefs about our relationship with reality and other minds? This is what I propose to verify. I will do so by adopting a common strategy; a thought experiment that evaluates the logical possibility of zombies in a possible world. In particular, I will assess what might happen in one of these possible worlds if a principle of accessibility, i.e. an epistemic window on a possible world, were applied.

2. *Possible worlds and accessibility criteria*

For the moment I assume that the conceivability (the logical possibility) of zombies is consistent. I also assume that a zombie universe is conceivable and logically consistent. The point, however understandable, is not whether zombies are logically consistent *per se*, but if there is room for maneuver regarding their conceivability, that there be sufficient room to enable us to somehow assess the epistemological consequences that would arise from the relationship of possible worlds to extensions. Imagine therefore that you have to introduce some intra-world accessibility condition that allows you to operate logically and epistemologically on counterfactuals[22]. In the case of Kripkean counterfactuals we know that the criterion of accessibility is given by the fact that the counterfactuals are possibilities nominally determined in relation to the real world. Expressions like "could have been otherwise" refers to well-designed possibilities that describe circumstances historically not realized due to items that historically existed or are still in existence.

[21] Primary intensions of "water", for example, are: "colorless liquid that fills the oceans and quenches" (Kripke's flaccid designators for instance). Then there are the "secondary intensions" as "H_2O", "$_{79}Au$" or "XYZ" (in a counterfactual), which form the *a posteriori* necessary identities. See: D.J. Chalmers, *The Conscious Mind*, cit., pp. 50-62.

[22] We usually talk about "operators" who select objects in different possible worlds. For a logical-metaphysical talk on operators see: D. Lewis, *Counterfactuals*, Blackwell Publishers, Malden (MA) 1973. A more intuitive way of thinking about accessibility to alternative worlds is described in Hilary Putnam's famous Twin Earth thought experiment (1975), that originally thinks to counterfactuals as to far away places that simply exist in the actual universe.

So what kind of criterion of accessibility is required to evaluate the concept of a zombie universe? One could accept a zombie universe merely as a type of possible maximal[23] universe where there isn't any consciousness and where organisms identical to us both physically and behaviorally could exist. I believe there would be nothing inconsistent if at some point we also imagined the creation of an epistemic bridge to such a possible world; in which case, we could then further extrapolate the events in this universe following the arrival of an astrophysicist with an "infallible" truth machine[24]. At the time of his arrival, the consciousness of the astrophysicist would be the only consciousness in a so-conceived universe. Hypothetically, he knows he is in a zombie universe but wants to see if what is experienced in his window is actually consistent and corresponds in some way to the truth and not to a contradiction. I am now in position to describe a possible interaction between a conscious individual and others, who by assumption are zombies. What might happen isn't hard to imagine; the astrophysicist could decide to study individuals of this universe by posing a few questions using the infallible truth machine.

The astrophysicist could devise the simplest of questions to assess the zombies' responses against that of the truth machine (T-M). The question addressed to the Z-Chalmers (i.e. the hypothetical Chalmers' zombie twin that the astrophysicist could potentially meet) is, "Are you a zombie?". From that question four possible scenarios emerge corresponding to the responses received from the zombie and the truth machine as shown below:

Zombie's answers	**Yes, I am a zombie**	**No, I am not a zombie**
T-M proves true	A	B
T-M proves false	¬A	¬B

An important premise must precede this examination; the zombie should conceivably interact with the astrophysicist. In this case it is essential for the zombie to understand the traveler's questions from the start. However, this assumption alone leads to the most problematic logical-phenomenological consequences for the conceivability argument.

[23] A "maximal universe" is a causally closed universe.

[24] Try to think of something similar to what is described in J.L. Halperin's *The Truth Machine: A Speculative Novel* (1996). An infallible truth machine is (at least) positively conceivable.

Case A

Z-Chalmers answers: "Yes, I'm a zombie." The machine checks and gives the output: "True."

How should we evaluate the first case? Might we have some reason not to believe Z-Chalmers? In particular, would we have any objective reason to contradict both the individual and the T-M in question, which hypothetically we know to be infallible? We know that Z-Chalmers is supposedly identical to Chalmers and speaks the same language of the traveler except for the fact that, according to the premises, Z-Chalmers has no conscious experience. The T-M, therefore, cannot fail the exam. Which is, at least at first, reasonable to believe if we accept the approach of the thought experiment, together with its consequences, thereby giving no objective reason to doubt that Z-Chalmers is (or would be) what he says to be. The problem is that the objective reasons brought into play by no means exhaust all the options. In particular there is the next logical problem; assuming that Z-Chalmers is telling the truth, according to the machine, how can we evaluate the concept of "truth" for Z-Chalmers. What, in other words, does the statement, "Yes, I'm a zombie" mean to Z-Chalmers? The semantic content of the statement corresponds to something like, "Yes, I am an individual identical to Chalmers, but devoid of conscious experience". The question is, could Z-Chalmers be engaged in individual observation around the content of this truth? I think the answer is clearly, no!

What is strictly required in order to allow Z-Chalmers make this kind of statement with a full sense of completion is that, in the least, 1) he assumes a phenomenal perspective on his own condition, and 2) the answer is willingly uttered, i.e. with the aim of stating something that he has semantically understood. Neither of the two conditions can be eliminated because the type of answer given properly concerns a fact of phenomenal consciousness, i.e. intentional recognition of his subjective condition. Z-Chalmers cannot satisfy the conditions 1 and 2. As a result, Z-Chalmers in principle cannot articulate meaningful utterances sensitive to the common concept of truth. The discrepancy between the content of truth and the hypothetical state of affairs are evidently in contradiction.

The only other counterfactual "experience" in which a similar situation could recur subject to the conditions on the absence of conscious experience is one in which Z-Chalmers would be a well thought out android robot (say R-Chalmers). But in this case we would state that R-Chalmers does not understand what it says and that it provides automated responses based on sophisticated software that controls its verbal and emotional responses.

Case B

Z-Chalmers says: "No, I am not a zombie." The machine confirms: "True."

How could this second case be evaluated? Are there objective reasons to doubt what Z-Chalmers says? Also in this case there seems to be no immediate objective reason for doubt. The machine is infallible as in the first case (with respect to the evaluation of the physiological responses of the individual) and we assume according to the principle of interpretative charity that Z-Chalmers is telling the truth. From the point of view of our basic epistemic commitments, there does not seem to be any conflict whatsoever. We are epistemically ready to recognize in the hypothetical condition of case B a situation in which the premise is false; that is, that it was in effect a zombie universe.

But what about the logical-phenomenological problem that emerges from this case? We reasonably have two possibilities. For the first we would have to admit the existence of such a counterfactual and the possibility of finding individuals identical to us (yet without conscious experience); however, the general premise is that this would not be the case, which is not a trivial problem. If the zombie tells the truth, then either there is sufficient reason to prove the premise or we deny the conclusion which is in contradiction with the premise. But why should we have to accept the premise that this would (in any case) be a zombie universe? I think someone could assess the following additional condition; true propositions expressed by Z-Chalmers overlook the epistemic condition, i.e. its (Z-Chalmer's) epistemic commitments (affirmation of the premise). This, however, would reasonably mean very little because in that case the traveler would be the only capable of semantically evaluating the statements and we would have the additional problem of defining the conditions of possibility for the evaluation of semantic content proffered by an individual of the zombie universe. If Z-Chalmers does not intend to express a minimum core of conditions that justify its utterances, the experiment is empty of meaning and not consistently proving that Z-Chalmers would be possible. The second case also seems to fall within the limited range of a well-programmed robot.

Case C

Z-Chalmers says, "Yes, I'm a zombie," but the machine proves the statement as "false."

In this third case there seems to be at least one objective reason to assert that the premise is not consistently qualified. We know that the truth machine is infallible and, based on this premise, we can admit that Z-Chalmers lied. But

what is the condition for which we would admit that Z-Chalmers would have lied in a so-described condition? The most reasonable solution, directly consequent to the introduction of the thought experiment, is that the physiological responses of Z-Chalmers would be somehow connected with the intention of providing an answer that has a content (true or false as it may be). Furthermore, one would say that this would confirm our intuition that the minimum necessary conditions are present for the utterance of true/false statements and that these conditions are related to the intentional character of subjectivity. As in the other two cases, this could also be a case in which the experiment would be misleading by virtue of the fact that Z-Chalmers could be a robot. All the same, this would not be the case in a zombie universe.

Case D

Z-Chalmers says, "No, I am not a zombie." The machine emits the output, "false."

Here there is an even more obvious immediate contradiction between the objective condition (the response of the machine and the zombie's response) which can be observed as neither objective nor subjective. But even if the response of the zombie were evaluated as an objective condition, it would have no effect on the experiment. The reason for this can be well illustrated. We know that the machine evaluates the physiological condition of the individual in question. But then the simplest assessment is that Z-Chalmers is somehow responsible for the utterance of a content of truth, all the same he knows he is lying, so the machine confirms a "false." Conclusion: either Z-Chalmers is not a zombie or he doesn't know he is lying; but then there is a logical conflict between the minimal epistemic conditions for expressing a rational content. (Remember here that one of the preconditions is that the zombie can provide answers because he is behaviorally identical to us and speaks our language.) It seems that the premise of the experiment tends, in all four cases, to invalidate the idea that Z-Chalmers is logically consistent, which is what I proposed to assess rigorously.

Even in the latter case there is the spectre of the robot. A robot could be so well-designed that it could falsify the experiment. In case D, the robot could have given an emotional response that falsifies the experiment, but this would only be admissible regardless of further logical evaluations that we would feel the need to put into play. We know that the case of the robot is not admissible, and any other case seems to disconfirm the conceivability of the zombie.

3. Intentionality, epistemic commitments and semantic contents

At this point, some or perhaps all the supporters of the zombie conceivability may argue that this mental experiment dramatically forces the assumption of logical conceivability. In particular it may be asserted that it is unnecessary to evaluate the boundary conditions of the statement that acknowledges the existence of "individuals identical to us and unconscious" (formally "P∧¬Q")[25]. If "P∧¬Q" is assumed logically consistent, then it is metaphysically possible. However, we could argue that the mere assertion of consistency is very little to lay on the table in order to state, while denying the contradictions, that zombies are logically consistent. The epistemic conditions are inevitably raised, from the semantic interpretation of "P∧¬Q," i.e. by the idea itself expressed in "individuals identical to us and unconscious."

The question just posed cannot be bypassed without evaluating, with an epistemic window, the implications intuitively raised from the main premise. This can be put another way, reiterating what Kripke admits about the metaphysical consequences of a purely epistemological evaluation. It is clear that the epistemic conditions inevitably emerge from the semantic interpretation of "P∧¬Q," i.e. from the idea itself expressed in "individuals identical to us and unconscious." Scientifically significant discoveries of physics or other sciences, whose domain of objects is given by the set of properties closely supervening to those of physics, have metaphysical and not trivial consequences; thus it is in the already described case of the discovery that the "internal" structure of water is H_2O. This discovery necessarily fixes *a posteriori* the reference, in the worlds in which there is H_2O, and we will say that, "water is H_2O."

Similarly, the epistemic window open in this experiment, although there are no such similar conditions (in the sense that we do not have zombies at hand), is needed to enable us to evaluate the minimum conditions of the intuition that there may be some "P∧¬Q." Such an epistemic window should provide room for maneuver to evaluate the consistency of zombies. But it seems that such a window gives rise to issues generated from the evaluation of the logical-phenomenological radical meaning of subjectivity, i.e. the minimum core that conditions the conceivability of situations in which it would make sense (i.e. would be consistent) to evaluate the utterances deemed rational for an individual "P∧¬Q." The epistemic window suggested basically coincides with the field of meaning available when interacting with other people. In ad-

[25] Cf. D.J. Chalmers, *The Character of Consciousness*, Oxford University Press, New York 2010.

dition, we believe (at least I do) that the assessment of the content of meaning presupposes the epistemic responsibility on the contents themselves. We will simply say that it is logically consistent that a person identical to us can take epistemic responsibility for what they state, unless one of the following conditions is not satisfied (separately), 1) the individual has a point of view which reflects what they say, and in that case can affirm a true or false response, or 2) the individual says things that are meaningless for them as they are devoid of consciousness (although the given statement might make sense to us, i.e. to conscious beings), or 3) the case is fallacious because we are dealing with an impostor, or robot in this case.

It is clear that the second case is the most controversial. However, it is difficult to imagine a case where a zombie could coherently articulate meaningful sentences, intentionally designed to express a condition based on the evaluation of a specific question, apart from taking some of the rationally accepted core epistemic commitments, which determine a kind of individual responsibility in cognitive and rational terms. In other words, it is difficult to accept the zombie's answers and take an epistemic commitment such as, "everything is dark and silent within." This seems to go against our most basic intuitions or, if you like, against the minimum evaluative (rational) schemas constructed on the base of our cognitive system. Given these difficulties, my conclusion is that zombies are logically impossible, because zombies do not pass such a test of consistency and truth.

4. *Other epistemic windows*

In his article (1999)[26], Robert Kirk developed a kind of experiment to evaluate the logical conceivability of zombies, and his conclusions do not differ much from the present method although his experiment presents a fallacy that I would now like to comment on. Kirk's experiment bets everything on the implications that result from a window of evaluation based on a cognitive situation. The experiment is simple and there is no need to fully reconstruct the logical setting of the premises of the experiment itself. Kirk evaluates the fundamental point expressed by those defending the conceivability of zombies, which is represented by the Cartesian idea that we are not identical to our body; hence, not being identical to the body in their view means, "I consist of

[26] R. Kirk, *Why There Couldn't Be Zombies*, cit.

my body, my 'zombie companion,' we might call it, combined with my non-physical qualia."

Kirk introduces his objection at this point asking, "Can this compound tell the difference between, for example, the smell of tea and the smell of coffee? Of course it can discriminate the two smells – my zombie twin can do that - but this is not the point. The question is: can it tell the difference between the 'subjective character' of the smells, between their qualia?" This further opens, so to speak, the epistemic window. Kirk comes to the following conclusions and says, "Being able to tell the difference in that sense requires being able to detect to alleged non-physical qualia in the first place: and that requires, as an absolute minimum, being sensitive to, or affected by, them. We already know the zombie is insensitive to anything non-physical, so cannot detect non-physical qualia or differences between them. The next thing to notice is that qualia cannot detect such differences either. Qualia are, at best, that between which there are differences to be detected; they cannot also be that which does the detecting."[27]

The point is clear enough. Kirk raises the question that we are not just the sum of a physical body plus the qualia. This is quite evident from the idea which keeps the two components separated as a premise and connects them, suddenly, in order to assess what (in principle) could happen. Kirk's epistemic window seems to necessitate the recomposition of the individual and presupposes the capability of the body. Then, to Kirk's conclusion we have to consider Chalmers' theoretical position on supervenience[28] which is also attributed to intentionality, yet not to the ability to have phenomenal experiences of any type in order to assess the qualia. But the premise is that the zombie cannot assess non-physical elements by comparing them with each other in their specific intrinsic qualities; therefore, since it seems impossible to reconstruct the individual from the two separate sets of elements, one must reject the premise that, as they are usually defined, zombies are logically consistent.

The other point is that the qualia cannot influence the zombie to provide an independent evaluative window (as if the qualia to be evaluated were able to self-evaluate in some way and give zombies a package of differences or intrinsic values). The value of the differences is somehow rated by the subject and that, according to Kirk (and according to me), requires the ability to cognitively evaluate the so-given values. Otherwise, if the zombie were to judge the differences regardless of "sensitivity" to the differences, the case would clearly be nonsensical. (Returning for a moment to my point, what Kirk says is very reasonable,

[27] *Ibi*, p. 7.
[28] D. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory*, cit., pp. 74-75.

but what he does not say is even more reasonable; i.e. the intentional evaluative attitude/behavior is an inviolable core of subjectivity. In other words, my horizon of meaning is inherently evaluative from the simple sensation of "feeling myself", whatever that means. What I would like to emphasize in particular, also in the case I present, is the idea that if there are sufficient grounds to admit "as significant" the declarative sentences uttered by an individual, then the zombie conceivability suffers a formal defect, given *ab origine* by the idea of curbing ontologically only the qualia. So, from the point of view of a rational observer evaluating the specific case of "P∧¬Q," the idea of zombies should in any case be logically unsatisfactory.)

As a result, what does not work very well in Kirk's experiment is the fact of granting zombies and qualia the possibility to be reunited. We know of limit-cases in which it would be openly consistent to join together a partial zombie with his qualia. This is the case, for example, in individuals who have regained their sight. In such cases, the individuals do not recognize the difference between the perceived sensations such as colors. But in the long run, a blind man who has regained his sight will be able to cognitively evaluate the relevant differences, for example, between the sensation of the colors red and gray. The problem with Kirk's experiment is that it has a premise, recognized even by supporters of zombies, that we do not know what constitutes this "conjunction." Therefore, it cannot be *a priori* excluded that it is not sufficient to operate the mere separation of the various features of the human body and qualia, just as you cannot know in advance whether this operation of reverse reductionism suffers from an inherent fallacy[29].

5. *Intentionality and metaphysics*

The most controversial point of this discussion seems to rotate around the question of whether or not intentionality poses a separate ontological problem. Chalmers denies that intentionality poses such problems, but from the mental experiment proposed here it would appear to be a different situation. Intentionality seems properly qualified as the linking element that unifies the ontological and logical framework; and that does not allow the commitment to a logical consistency of zombies. This element, the horizon of the sense of subjectivity, reflects the radical character of the phenomenological framework

[29] Chalmers says much the same thing. See: http://consc.net/responses.html#kirk

which cannot be ignored in assessments that are often regarded as purely logical. This is actually the further argument I wanted to support through the mental experiment. If the intention of asserting a truth content were eliminated from the case "P∧¬Q," then we would have a consistent situation where an individual might respond coherently without any commitment to the discernment of the conditions under which A rather than B is stated. We know that it would be perfectly conceivable in the case of a robot, the robot states what it states because a software program is designed to provide particular answers.

Therefore, the case of the robot, which is not a real case of "P∧¬Q," is disregarded. It does not seem that there is any conceivable, consistent situation for which a "P∧¬Q" might have sufficient reason to make particular statements, regardless of certain minimal evaluations and regardless of the intention to provide some response that we would consider rational (i.e. as significant) in the situation described. All this seems, if I'm correct, to characterize the major difficulty for the conceivability thesis.

5.1. *Consciousness, unconscious and intentionality*

I will now discuss two possible objections to the arguments submitted. At this point of the discussion, the supporters of functionalism might object to what is said stating that I have not given any proof that the utterance of true or false sentences is strictly subordinated to the conscious assessment of intentional states, or phenomenal content. If one accepts the thesis of the multiple realizability of the mental and admits that intentional states are expressions of the type of physical system functionally capable of implementing them, then one can conclude that an artificially made Z-Chalmers' counterpart may genuinely manifest intentional behaviors when uttering the sentence "yes, I'm a zombie", without the necessity of satisfying the condition that Z-Chalmers must also be conscious of these intentional states. This still means that, by not having well characterized the distinction between a human and a robot, my arguments are exposed to the objection that if the functionalists are right, then Z-Chalmers could very well be a natural automaton or a silicon counterpart capable of implementing nontrivial intentional states on the basis of certain software, while being devoid of consciousness. If we assume that the intentional behavior is reducible to causal patterns of interaction between certain external events and physical brain processes, then we still have to admit - as any cognitive psychologist would - that behaviors which are coherent to the context could be implemented by a neural system thanks to an innate software of which we might have

no access through introspection (the modularity thesis[30]). This type of reasoning is usually shared by the defenders of the computational theory of mind.

These are strong objections, all the same it is nevertheless possible to present responses. First, I would like to point out that a behavior considered as "adequate" to the context cannot logically be considered an intentional behavior, in the same way that the software Word cannot be considered to have an intentional state as it responds to input that I send from the keyboard to write the following and what has been written so far. The phenomenal aspect of the letters in Times New Roman is not the aspectual form of intentional states of the Word software. Furthermore, to admit that certain behavior of a physical system identical to ours might not require consciousness profoundly underestimates our common and intuitive concept of "rationality".

We consider a behavior "rational" on the basis of which there is (or will be if we decide to implement it) a subjective responsibility, i.e. the subjective/conscious evaluation of intentional states with respect to a given context, and this is required for verbal utterances (true or false). But perhaps this definition does not help much. What, however, could help is what Searle discusses in *The Rediscovery of the Mind*[31]. I refer in particular to the "Connection Principle" (CP) and the concept of "aspectual shape" of intentional states (presented by Searle to answer the functionalist thesis) according to which intentional acts may be closely linked to the expression of unconscious brain events, i.e. processes of a computational nature that make it possible to exclude qualia from the framework of theories explicative of the mind, being characterized by purely syntactic-formal elements. Searle's attempt to include aspects of phenomenal consciousness in the context of the conditions necessary to characterize intentional states is what we are proposing.

First we have to characterize the CP. We admit that the intentional states of an individual, such as emotions, feelings and beliefs, are expressions of some type of neurophysiological activity. Searle's CP states that under normal conditions, any intentional state can be accessible to consciousness. This would be true for two main reasons. The first is that each intentional state, if genuinely intentional, has an aspectual shape defined by a particular type of intentional state for a particular individual who holds it in "private". In other words, if S believes that P, and P is a genuine intentional state, then P is a subjective value that manifests through its aspectual shape, i.e. the intrinsically subjective

[30] Jerry Fodor is the main defender of this theory. See J.A. Fodor, *The Modularity of Mind. An Essay on Faculty Psychology*, The MIT Press, Cambridge (MA) 1983.

[31] J.R. Searle, *The Rediscovery of the Mind*, The MIT Press, Cambridge (MA) 1994.

means by which P is believed by S. Second, if there are not sufficient reasons to state the contrary, in normal conditions each intentional state can be accessible to consciousness, because the brain is (according to Searle) that kind of physical system that is "able to produce" consciousness. In other words, a physical system, if it is a neurophysiological system identical to ours, in conditions of normal operation realizes aspectual shapes of phenomenal consciousness – save for the principal of causal closure. As stated by Searle, whereas unconscious intentional states are reducible to the physical states of the neurophysiological system, their intrinsic aspectual shape can be made explicit at a conscious level, thus dispensing the need to embrace some form of dualism. However, it's necessary to add another thesis to this one.

The aspectual shape of intentional states, according to Searle, is preserved (in the form of an objective value) in the neurological structure of individuals. This means that even the unconscious intentional states retain their aspectual shapes. From this, in principle, it follows that one day we may be able to identify the neurological aspects of an unconscious intentional state and thereby determine its aspectual shape (to establish the intentional state of a particular corresponding neuropsychological state). As I have understood Searle's theory, it can be said that the objective characteristics of an unconscious intentional state correspond to the dispositional-causal elements that may (i.e. causally) result in a given type of conscious experience. Since the unconscious intentional states have a physical nature that preserves an aspectual shape characterizable to the level of conscious mental states, it follows that consciousness is a causally determined product of our neurological system. For reasons of brevity I will pause at this point on Searle's theory.

Now it is clear that the concept of aspectual shape could be rejected along with the idea that an aspectual shape exists at an unconscious level, even if one is unable to bear reason on principle to deny that the concept itself could be valid and useful for science. Conversely, taking as valid the Connection Principle, what consequences can be drawn for answering the functionalist's objections to the main thesis of this work? I will not repeat Searle's answer to the zombie case[32], which relies on Quine's indeterminacy of the translation argument, because I think that the Connection Principle and, in general, the doctrine of biological naturalism supported by Searle is an extraordinary opportunity to draw some radical conclusions about what non-reductionist materialism can offer in response to the functionalists.

[32] *Ibi*, p. 163.

I have adopted one a Searle's analogies that I find very powerful, which features a kind of physicalism in which consciousness finds its natural place. Kripke's lesson on essential properties, intended as the "Individuation Principle," will also be kept in mind. The analogy compares consciousness as an aspect of the human neurological system in the same way that being liquid is an aspect of water. In both cases, it is about physical aspects reducible in terms of definition and not in phenomenal terms. What does this analogy allude to? It reveals something that Searle explicitly states elsewhere in the text, namely that consciousness is caused by the elements characterizing the neurophysiological system according to a micro-macro or low-high ratio[33]. So, starting with zombies, the premise of the entire discussion is that a zombie is a physical system identical to me, i.e. identical "molecule by molecule" and for all those material aspects reducible to the basic properties of physics; in addition, his neural system works exactly like mine. It follows that if Connection Principle is valid and the reader also keeps in mind Occam's Razor[34], then in the situation I have described, it is "impossible" for the Zombie in the mental experiment to give an answer (that is an expression of genuine intentional states) without a neuronal system which brings up to a conscious level the aspectual shapes that characterize the semantic contents of utterances heard from the astrophysicist or those uttered by the zombie itself. In other words, if the zombie possesses genuine intentional states then, given the identity, even its states must have aspectual shapes (and provided that the identity is valid for all sets of physical properties causally supervenient to basic ones) which must necessarily manifest themselves as characters of consciousness. All this challenges Cartesian argumentation.

This reasoning contradicts the logical possibility of zombies, relying simply on the expression, "individual just like me," a phrase that (if considered on the basis of Connection Principle) excludes the other part of the conjunction, i.e. "unconscious". From now on, formally as well as metaphysically, we would say that the conjunction "$P \wedge \neg Q$" is a contradiction in all the possible worlds. From this reasoning, would we perhaps be able to derive a regression of the zombie to a mere "potency" as in the Aristotelian sense? In other words, one could probably say that a zombie is such only to the point that it is non-thinking, i.e. only until a kind of neurophysiological activity takes place that, based on the very definition of "zombie", must be identical to that of normal people. But it is clear that we do not do anything in a redefinition of "zombie" as such,

[33] *Ibi*, pp. 124 ss.

[34] Keep in mind that, with the same physical condition is theoretically more expensive (in explanative terms) to support the idea that my identical twin would be unconscious.

and that, in any case, the answer would be negative because intentional states of the individual in question should have aspectual shapes even at the unconscious level (according to the reasoning).

At this point, one can argue that it is not at all obvious that intentional states should have aspectual shapes, especially if such attitudes are formal expressions, namely logical strings of symbols computed by software "blind" to semantic contents. I believe that this argument is irrelevant for two reasons. The first is that under normal circumstances intentional states have an aspectual shape that determines the first-person ontology, and this is a fact. The second reason is that even if such computational systems could be implemented by silicon counterparts, the corresponding computations (having no aspectual shape) could not be considered genuine intentional states. To explain it as Searle, the behavior of such counterparts would be judged "as if" they were rational agents[35], but the truth is that their behavior would be mechanical like that of much simpler machines, like a hydraulic system through which water flows.

This reasoning may be carried even further. Remember the analogy between the brain and consciousness and the liquid appearance of water. The phenomenal aspect of physical bodies (such as being liquid as in water or yellow as in gold) is causally supervenient for certain thresholds of realizability according to their microphysical properties. This calls into question the thesis of the multiple realizability of the mental aspect. In fact, we say that aspects from water such as being liquid, boiling at 90-100°C, as well as being colorless and odorless cannot be created using gold, silver or any other element other than oxygen and hydrogen and possessing the same boundary conditions. It's also necessary to bear in mind that these aspects of water are not mere primary intensions, but features causally determined in a causally closed universe. At this point I wish to push beyond what Searle asserts in his work. If what I am stating is true, then it is very likely that a structure of silicon (i.e. a counterpart of any type of substance other than that of which I am actually formed) cannot be implemented by or capable of consciousness. Therefore, if Searle is right and his type of doctrine is the best defense that we have at our disposition, then it is very reasonable to assert that there cannot be counterparts of human beings

[35] Once again, the reasoning might be rejected stating that Searle does not provide sufficient and necessary criteria to establish that the intentional states – having aspectual shapes – are the only true intentional states. However, at least intuitively speaking, Searle's argument is always better than the alternative, namely the idea that the material dispositions of an artificial hardware can implement truly intentional software.

that are equipped with software capable of operating on states which we would define as genuinely intentional.

In other words, counterparts would act plausibly "as if" they were able to understand a biological speaker's verbal utterances but in the absence of any real form of intentionality and thus without any reference to content. The software, or set of formal rules which they could have been given, would thus be behavioral dispositions, rules to allow mechanical response to environmental stresses, and not intentional states with aspectual shapes. So again, this too is to be considered unsatisfactory in regards to the idea of a zombie in its standard definition.

The concept that the material substrate is likely to significantly affect the feasibility of truly intelligent (e.g. intentional) systems is not generally considered problematic by functionalists. Perhaps this comes from the fact that reductionists often naively compare the computerized processes of a machine to cerebral processes without paying attention to the intrinsic value and determinates that distinguish organic systems from inorganic ones; it is worth assessing the radical consequences of this version of materialism if for no other reason than the fact that no one up to now has succeeded in artificially producing genuine conscious systems, namely possessing intentional states. From an epistemological point of view, it does not seem at all reasonable to discard *a priori* such a possibility, arguing that thought experiments such as that of Putnam's Twin Earth[36] are *de facto* a demonstration of the genuineness of the type of reductionism defended by functionalists.

It is correct to take note that in the original version of the experiment, Twin Earth is not conceived as a counterfactual, but as a distant place in the known universe. However, even if one were to hold Twin Earth as a logical possibility, an obvious fallacy would be encountered. This fallacy is seen by observing that in this universe, bearing in mind the causal closure, it is not possible for the water on Twin Earth (XYZ)[37] to have a microscopic structure completely different from the water (H_2O) on Earth, subject to all the boundary causal interactions, such as phenomenal aspects and causal interactions with other biological systems (considered physically identical to those present on Earth). For example, Putnam suggests that on Twin Earth, water "tastes like water and it quenches thirst like water,"[38] despite having a completely differ-

[36] H. Putnam, *The meaning of "meaning"*, first published in K. Gunderson (ed.), *Language, Mind and Knowledge* (Minnesota Studies in the Philosophy of Science, n. 7), University of Minnesota Press, Minneapolis 1975, pp. 131-193, University of Minnesota Press, Minneapolis 1975.

[37] "XYZ" is the hypothetical chemical formula for the "watery stuff" on Twin Earth.

[38] H. Putnam, *The meaning of "meaning"*, cit., p. 140.

ent microstructure. Personally I don't know whether there are liquids other than water that quench in the specific sense of the term. There are many water-based compounds such as beer or wine, but even in these cases the function of quenching thirst is accomplished by the H_2O molecules, which is not trivial from either an epistemological or metaphysical point of view when deciding to set a thought experiment. To have a liquid analogous to water (H_2O) for the assortment of truly relevant aspects for a physicalist analysis, we would have to conceive a universe completely different in its microstructure. But that would not be of any help for a defense of functionalism, because then we would not have any kind of epistemic access to relevant facts. These are my responses to some possible objections.

Conclusions

In evaluating Chalmers' conceivability thesis, the very act of posing questions to an entity identical to us but unconscious serves to avoid all the weaknesses found when there is a commitment to assess the coherence of the mere concept of zombies. Inconsistencies also show up in the case of Kirk, who temporarily violates the requirement that "P∧¬Q" be logically consistent. In so doing, Kirk violates to all effects the premise, going from "P∧¬Q" to "P∧Q." By contrast, an argument such as the one I have proposed aims to keep the discussion around logical data from the premises of the topic, thus deriving from the main thesis the most important epistemological consequences. The utterances proffered by the hypothetical zombie, in each of the four cases, appear to violate both a logical and epistemological element that cannot be eliminated.

If "P∧¬Q" is consistent, then there is no reason not to keep an epistemic window open in order to derive some immediate logical consequences on the behavior of a "so and so" individual. All the same, such logical consequences seem dramatically inconsistent with certain radical elements of our phenomenology, which is the foundation of the conceivability of anything. These elements would need to be considered basic so that "P∧¬Q" would remain consistent. As noted several times, the very possibility that sufficient reasons would be given to enable belief that the behavior of a zombie would be consistent also depends on the values of truth and semantic contents, in other words, what such an individual could reasonably utter. Additionally, what a zombie could most likely never do, is to utter significant sentences on its condition or internal state as a zombie.

Similarly, if one decides to support the idea that a counterpart of silicon could be able to manifest intentional states, it is important to remember that

there are some theoretical difficulties which functionalism will inevitably come up against. This alone would seem sufficient argument for the idea that a zombie is not logically consistent, at least not consistent in the strong version as presented by Chalmers, where only the qualia are irreducible to the explanations of physicalism.

Abstract

The subject of zombies is one of the most discussed and controversial topics of philosophy of mind. In this paper I will first examine the main argument of zombies, providing a summary of the current discussion. Then I will introduce a thought experiment, an epistemic window on a metaphysical scenario. By the thought experiment I will argue that zombies are logically impossible. Further I will discuss another recent epistemic window. Finally I will provide some other logical consideration to prove that intentionality is not reducible to the cognitive functional aspects of the mind and that, moreover, the subjective recognition of semantic contents is necessary in order to consider as sensical the verbal behavior of a zombie.