DRAFT! (15 MARCH 2004)

CHAPTER IV

CHALMERS'S ZOMBIE ARGUMENT

Conceivability arguments: from substance-dualism to property dualism.
Phenomenal concepts: denying the conceivability-possibility link.
Conceivability and the epistemic/ontic divide: two-dimensional semantics.
The inconceivability of zombies.*

***Problem 3*: Why zombies seem conceivable?**

In this chapter I will discuss the argument from the conceivability of zombies, first proposed by Robert Kirk (1974), and made more popular and widely discussed by David Chalmers in his book, *The Conscious Mind. In Search of a Fundamental Theory*. The argument goes something like this. If physicalism is true, then there cannot be a world that is a physical duplicate of ours (that is, where everything is physically like in our world), which is not a duplicate *simpliciter* of our world (that is, which does not contain anything more or less than what our world contains). But zombies are conceivable: creatures that are physically exactly like us, but which creatures lack conscious experiences. Therefore, physicalism is false. I will first briefly discuss this sort of argument

in a historical context, then I will discuss some responses to it, and finally, I will try to show that it is unsound.

4.1 CONCEIVABILITY ARGUMENTS: FROM SUBSTANCE-DUALISM TO PROPERTY-DUALISM.

The recipe for a conceivability argument is something like this. First, you make a claim about what is conceivable. Then you infer what is possible and necessary. Finally, you conclude with a claim about how things are or have to be in the actual world. Conceivability arguments are sometimes called 'epistemic' or 'modal' arguments. Kripke's argument, which we considered in the second chapter, is also a conceivability argument, as it begins with (1) the epistemic claim about mental/neural separability, then (2) makes a modal claim about necessity and possibility, and finally (3) a claim about the falsity of materialism (something about the actual world).

Historically, the use of such arguments goes back at least to G.W.F. Leibniz, who exposed the famous mill argument or thought-experiment. In section 17 of his *Monadology* he wrote:

> "Moreover, it must be confessed that (3) perception and that which depends upon it are inexplicable on mechanical grounds, that is to say, by means of figures and motions. And (1) supposing there were a machine, so constructed as to think, feel, and have perception, it might be conceived as increased in

> size, while keeping the same proportions, so that one might go into it as into a mill. That being so, (we should, on examining its interior, find only parts which work one upon another, and never anything by which to explain a perception. Thus it is in a simple substance, and not in a compound or in a machine, that perception must be sought for. Further, nothing but this (namely, perceptions and their changes) can be found in a simple substance. It is also in this alone that all the internal activities of simple substances can consist." (*My emphasis and numbering, I. A. A.*)

Leibniz's argument is more or less following our recipe. The only thing that is missing, or at least it's not made explicit, is the modal claim about possibility, namely, Leibniz did not feel the need to state a condition of the following sort:

(C-P) *If something P is conceivable, then P is possible (it is not necessary that non-P).*

Another observation I want to make is related not to the form of the argument, but its conclusion's content. It is an argument for *perception* as being simple and having to reside in some *simple substance*. Perception here can, I think, be understood as phenomenal feel, as quale. This point is important in what follows, as the idea of the need to state the existence of a simple (mental) substance is indeed tightly related to this form of argument, where the issue of the truth of C-P is either not considered important, as here in Leibniz's argument, or taken for granted.

A similar argument, but this time clearly for the existence of two separate substances, was put forth by René Descartes. He argued from the conceivability of disembodiment, i.e. of his conceivable existence without a body, to the separateness of mind and body.

There are some important elements that we should enumerate here, regarding the early conceivability arguments.

a. The idea of soul, mind, or perception, as simple, unitary, indivisible *versus* the idea of matter as divisible, having detachable parts, moving mechanistically.
b. The idea of the disparateness of these two kinds of existence, and hence inexplicability of the mental in terms of the material. (This is the same as the epistemic premise of conceivability)
c. The move from these ideas to the modal separability of the mental and the material (or mechanistic).[1]

This sort of conceivability arguments came to be heavily contested, especially after Kripke's ideas regarding a posteriori necessity. As I pointed out in Chapter II, one of Kripke's novel contributions was the idea of necessity a posteriori. It is relevant here, as it is this idea that was mainly used to attack the early, standard conceivability arguments as regards the mind/body relation. The basic

[1] There is a difference here between Descartes and Leibniz's approach, as the former states indeed constraints on conceivability, such as to make it a guide to possibility, and finally to a conclusion about actual states of affairs. As we saw, Laibniz is content with the simple fact that one cannot explain qualia by simply observing the internal mechanics of the mill. Descartes, on the other hand, makes explicit a kind of important constraint on conceiving, as he talks about 'clear and distinct perceiving', from which he infers possibility.

point of criticism against these arguments was that it may well be that it is conceivable, for instance, that one is disembodied, but that does not rule out the possibility that one is having a modal illusion, just as one initially has when, for instance, thinks that there could be water without H2O. Then, it may well be the case that we can conceive of our disembodiment, but it hasn't been shown that this conceiving is reliable: it may be that there is an a posteriori necessary connection between mind and body.

Take the classical example 'The Morning Star is The Evening Star'. We can conceive that this is not true, that is, that there are worlds where these correspond to two stars that are distinct. In other words, the two concepts, or, better, names, are separable; they are not linked a priori. But then rigidifying them will show us that there couldn't be such a case. They are identical and necessarily so. It is only our conception of them that was inadequate, so as for us to make the modal mistake of thinking they could have been non-identical.

Similarly with me and my body: it may well be a modal illusion my ability to conceive of my disembodiment. Moreover, I don't have a very precise idea of 'me', to be honest, and so there may be a necessary connection between me and my body, even if prima facie there doesn't seem to be one.

This attack on the early conceivability arguments is I think basically correct. But this brings us to the more recent conceivability arguments in the philosophy of mind. They are arguments for property-dualism, rather than the classical disembodiment-based substance-dualism. All the arguments I briefly presented in Chapter I have been mounted against physicalism and for the conclusion of a dualism of properties. The main difference between the two kinds of argument –for substance versus for property-dualism—lies, in my opinion, in the much weaker premises a property-dualist conceivability

argument needs. A property-dualist accepts that all substances are physical, but he will insist that there are physical substances (objects) which have some properties other than physical ones, and most importantly not reducible or identifiable with physical properties.

I will focus here on David Chalmers's zombie argument, as that is one of the more vivid ways to expose the problem for physicalism[2]. A zombie is a molecule-for-molecule duplicate of a conscious person, say, me, but it lacks conscious experiences. It behaves just like me[3], reacts to various stimuli in exactly the same way, makes the same verbal reports as I make, etc. Yet it lacks the qualia that are instantiated by my experiences. There is nothing it is like to be a zombie.

Is this scenario conceivable, non-contradictory, or logically possible? For the moment, I think we should accept that prima facie it is. It is useful to quote Chalmers at this point (1996, p. 96):

> "I confess that the logical possibility of zombies seems equally obvious to me (*as that of a mile-high unicycle, I. A. A.*). A zombie is just something physically identical to me, but which has no conscious experience – all is dark inside.[…] I can discern no contradiction in the description. <u>In some ways an assertion of this logical possibility comes down to a brute intuition, but no more so than with the unicycle.</u>"

[2] It is a bit improper to call Chalmers's main contribution to the debate 'the zombie argument', as that argument is just one of the epistemic arguments he discusses; the main contribution of his is, I think, the 2D-semantics based attack on a posteriori physicalism, which we will touch upon in section 4.3.

[3] Since it is a physical duplicate of me, then it will certainly be a functional duplicate as well, as functional concepts/properties are deducible, follow a priori from physical ones.

Chalmers is right: we have no special or general analytical method for finding out whether something is logically possible, but most of the times we have a brute intuition in this sense. Yet I think the precise issue here, that of zombies, indicates only of prima facie logical possibility. But first let me briefly introduce three kinds of possibility, in order for our discussion of the argument to lay on some theoretical ground regarding modality. The three kinds of possibility are: logical, natural, and metaphysical.

Logical possibility is the weakest, most liberal kind of possibility, and is best thought of as non-contradictoriness of the described scenario. Conversely, logical necessity is the strongest kind of necessity. Logical possibility is I think the most important notion in philosophy, as it plays a key role in thought experiments and counterexamples to putative conceptual analyses of notions.

Natural possibility is tighter than the logical one, in the sense that the set of naturally possible scenarios forms a subset of the set of logically possible ones. For example, a 100 tons homogeneous piece of Uranium is naturally impossible, but it is logically possible; I, at least, can conceive of it[4].

Metaphysical possibility is a trickier notion. I myself think that it is the same as broadly logical possibility. A certain scenario is metaphysically possible if it is logically possible in the broad sense of 'logical'. To see what this broad sense is, it is useful to consider a stock example from Kripke – the scenario in which water is not H_2O. What can we say about this scenario in terms of the two

[4] There are very interesting issues here. There are philosophers who argue that laws of nature are necessary, and so, they would say, the 100 tons piece of Uranium is not just naturally impossible, but also logically. I think this is false. Even if laws of nature are necessary, this won't change much. We can still conceive a world where there is something exactly like Uranium in all respects, except that it doesn't disintegrate even in such a huge quantity. We may

clearer notions briefly explained above? It is not naturally possible, because water just is H_2O in the actual world. In fact, laws of nature are not even relevant to the issue[5]. Is it logically possible? Some would say here that it is logically or conceptually possible, but not metaphysically so[6]. I would say that to the degree that it is possible or impossible, it is broadly logically so. In other words, I wouldn't subscribe to the view that we have a separate category of metaphysical possibility and one of conceptual possibility; all possibility is conceptual, but it depends on how we approach our concepts. In our specific case, I would say water without H2O is conceivable and conceptually (=metaphysically) possible under the interpretation of 'water' which takes it as a non-rigid description, and it is not conceptually (=metaphysically) possible under the interpretation of 'water' which takes it as a rigid. More on these issues in the third section of this chapter.

Let us now return to the argument from the conceivability of zombies. The argument is based on the premise, accepted by everybody in the debate, that if zombies are metaphysically possible, then physicalism is false. The argument is, according to my understanding[7], the following:

---

call it Uranium*, this is not relevant for the metaphysics of modality, but rather involves purely linguistic considerations.

[5] Thus I agree here with Kit Fine, who argues in an excellent paper (2002) that we should take the notions of natural and other types of necessity and possibility as independent from each other, i.e. not reducible to either of them.

[6] For example Stephen Yablo (1999, 2002) strongly believes that cases like this are cases of what he calls 'conceptual possibility' without (a reliable route to) what he takes to be 'metaphysical possibility'. Yablo bases his views on Kripke's examples of a posteriori necessity. I think Yablo's distinction is not justified, as I think Kripke's examples are more rationally accounted for by the Jackson-Chalmers 2D interpretation of a posteriori necessity. See section 3 of this chapter for my preferred interpretation.

[7] I think my formalisation of the argument is the best for the purpose of thoroughly discussing the issues, but without any special knowledge of semantic or other theories. Alternative formulations seem to me either too complicated (Chalmers 1999, 2002; Andrew Melnyk 2001) or "too introductory" (Crane 2001).

1. If zombies are logically possible, then zombies are metaphysically possible.
2. If zombies are metaphysically possible, then physicalism is false.
3. Zombies are conceivable.
4. If zombies are conceivable, then zombies are logically possible.
5. Zombies are logically possible. (from 3 and 4, MP)
6. Zombies are metaphysically possible. (from 1 and 5, MP)
7. Physicalism is false. (from 2 and 6, MP)

I will now turn to the a posteriori materialist response to this argument. Again, I think, given my way to formulate the argument, we can distinguish a more standard (and less interesting) from a subtler a posteriori materialist reply. I won't discuss them in separate sections.

4.2 PHENOMENAL CONCEPTS: DENYING THE CONCEIVABILITY-POSSIBILITY LINK.

The more standard a posteriori materialist reply to this argument is to deny the first premise (e.g. Michael Tye 1985, Brian Loar 1990/1997, Papineau 2002, John Perry 2001), while the subtler one is to deny premise 4. (Yablo 1999). I think we can also make a distinction within the first category, between a less and a more elaborate reply. Let us take them in an order from less to more subtlety and sophistication.

The early such replies were no different from what I earlier touched upon regarding the attack on the early conceivability arguments, which were concerned with proving substance-dualism. An example is Tye (1985), who, after exposing a quite complex and interesting intentionalist theory of consciousness and of qualia in particular –which he also takes to be a physicalist theory—arrives at the problem of zombies, and responds by simply appealing to the classical Kripkean examples of necessity a posteriori – which he takes to be examples of "conceivability without possibility". The classical examples of a posteriori necessity are, as I have already clarified, at most counterexamples to the move from conceivability to possibility of disembodiment or the separation of mind (person) from body. Second, as we will see in the next section, the Kripkean examples are not especially usable by the materialist, as they are perfectly well accounted for by the 2D semantics, which semantics can at the same time be used as a rationalization of anti-materialism. Let us then turn to a more sophisticated position, that of Loar (1990/1997), followed with practically no improvement upon by Papineau (1993, 2002) and Katalin Balog (1999), among others.

Loar (1990) argued that besides the case of co-reference of two distinct predicates to an object, we should also recognize co-reference of two distinct predicates to a property. In this way, one can use the same sort of considerations against the epistemic arguments for property-dualism as one would use against substance-dualism, namely the opaque co-reference of these predicates. Loar takes the main fallacy of the qualia objections to physicalism to be what he calls the "semantic premise", and formulates it as follows:

Semantic Premise = A statement of property identity that links conceptually independent concepts is true only if at least one concept picks out the property it refers to by connoting a contingent property of that property. (*My emphasis, I. A. A.*)

Loar's defense of physicalism consists in denying this premise, and then offering a view of phenomenal concepts that makes them special in such a way that physicalism in it's a posteriori version is saved, according to him. Let us analyze the premise first, and then Loar's theory of phenomenal concepts.

I have three short critical points regarding Loar's approach to the premise. First, the second phrase I have emphasized is hard to understand otherwise than as referring to concepts separate in a way opposed to "non-conceptually" or "metaphysically" separate ones. But what could "metaphysical separation" *of concepts* mean? I don't see. The issue turns on Loar's understanding of modality, namely on the distinction he draws between metaphysical and conceptual (a priori) possibility/necessity/contingency. This brings us to the next point.

As regards the third phrase I have emphasized, Loar does not offer any theory of contingency (or modality in general), so everything turns on the *deus ex machina* of the notion of metaphysical modality as opposed to the conceptual one. My argument against the a posteriori materialist response to Kripke's argument in Chapter II was based on a clear notion of contingency, that of conceivable situations (worlds). Loar cannot appeal to such an understanding, since it would buttress the *Semantic Premise*. At the same time, he does not offer any alternative understanding of contingency. A more elaborated criticism

of metaphysical as opposed to conceptual modality will be offered in the next section.

Third, if one is prepared to deny the Semantic Premise, why shouldn't he be prepared to deny a similar principle in the case of object identity statements? For example, one would then have to deny its application to this sentence:

"The Morning Star = The Evening Star"

This would mean nothing else but to go against the background that made possible Kripke's points regarding a posteriori necessity to be so widely accepted; and Kripke's theory of a posteriori necessity is certainly the basis for a posteriori materialism, including Loar's approach, though we should keep in mind that his version does not rely on the standard Kripkean examples of a posteriori necessity, as we shall now see, as we turn to his toy theory of phenomenal concepts. As a preliminary conclusion, Loar's denial of *SP* is almost like a theoretical self-denial.

Loar's toy theory of phenomenal concepts is, in a nutshell, the following. They are a special sort of what he calls 'recognitional concepts' (p ??):

> "They have the form 'x is one of that kind'; they are type-demonstratives. These type- demonstratives are grounded in dispositions to classify, by way of perceptual discriminations, certain objects, events, situations."

There is no problem I think with understanding what recognitional concepts are: they are concepts that we unscientifically, heuristically, and without a precise

description use mainly when we perceptually discriminate among some types of things, this is why they are called type-demonstratives. Further, as regards phenomenal concepts, they are special in a way (p ??):

> "Phenomenal concepts are recognitional concepts that pick out certain internal properties; these are physical-functional properties of the brain. They are the concepts we deploy in our phenomenological reflections; and there is no good philosophical reason to deny that, odd though it may sound, the properties these conceptions phenomenologically reveal are physical-functional properties -- but not of course under physical-functional descriptions."

Though I have some trouble with understanding phenomenological revelation of physical-functional properties, I will consider that it is clear. Let us see how Loar explains the possibility to rebut the *Semantic Premise* and then how he thinks he solves the apparent problem for materialism, namely that it cannot accommodate phenomenal concepts, since these will refer to phenomenal properties distinct from physical ones. First, how he explains (p. ??):

> "Rebutting the semantic premise of the knowledge argument requires making sense of the idea that phenomenal concepts conceive physical-functional properties 'directly', i.e. not by way of contingent modes of presentation."

As regards this idea, I offered a clear refutation of it in Chapter II, so I won't rehearse that criticism here, but offer another argument, which has not been put forth so far in the literature, and I think it may be at least as good as the one I proposed in Chapter II. Let us then see how he thinks he solves the problem mentioned above and then I turn to criticism:

> "The physicalist thesis implies that the judgments "the state a feels like that" and "the state a has physical-functional property P" can have the same truth condition even though their joint truth or falsity can be known only a posteriori. I mean, same-condition-of-truth- in-a-possible-world. For truth conditions are determined in part by the possible world satisfaction conditions of predicates; and if a phenomenal predicate directly refers to a physical property, that property constitutes its satisfaction condition. […]
>
> Even on the anti-physicalist view, phenomenal concepts are recognitional concepts, and we have 'direct' recognitional conceptions of phenomenal qualities, i.e. conceptions unmediated by contingent modes of presentation. Evidently it would be absurd to insist that the anti-physicalist hold that we conceive of a phenomenal quality of one kind via a phenomenal mode of presentation of a distinct kind. And why should the physicalist not agree that phenomenal recognitional concepts are structured in whatever simple way the anti-physicalist requires? That is after all the intuitive

> situation, and the physicalist simply claims that the intuitive facts about phenomenal qualities are compatible with physicalism. The physicalist makes the additional claim that the phenomenal quality thus directly conceived is a physical-functional property." [*My emphasis, I. A. A.*]

My criticism is the following. A physical-functional property is a relational and dispositional one[8]. If a physical-functional concept refers directly to a physical-functional property, then it refers transparently to it, as that is just the meaning of 'directly'. Physical-functional concepts refer directly to physical-functional properties: this is agreed by all physicalists[9]. A phenomenal concept is not a relational and dispositional concept, but a monadic and categorical concept. If it refers directly to a phenomenal property then it does it transparently, as that is just the meaning of 'directly'. So let us then see what the case of opaque co-reference of physical-functional and phenomenal concepts to one and the same physical-functional property makes us to be committed to. Suppose we have two concepts each belonging to different kinds of concepts: C1 (a phenomenal/monadic/categorical concept) and C2 (a physical-functional/relational/dispositional). Then suppose we have a physical-functional

[8] A functional property is certainly causal-dispositional. On the other hand, there are people who think that physical properties are not dispositional; for example those whom George Bealer (2002) calls 'right-wing materialists', that is, materialists who think that there is some physical essence that escapes a causal-dispositional analysis, and equate qualia with this essence. I think there is no such essence, and if it were it would buttress the doctrine of type F monism (Chalmers 2002), which consists in positing a ('physical') reality underlying the reality described by physics (which makes use of dispositional notions). I think that is not physicalism, though see Daniel Stoljar (2001) for an attempt at giving a physicalist interpretation for this doctrine. [Parenthetically, Bealer criticises Chalmers in that his 2D argument is not effective against these right-wing materialists; I think Bealer misunderstands or misinterprets Chalmers]

[9] Otherwise they would have to believe in a reality (possibly menatal) deeper than the reality described by physics.

property *F*. Finally, we have C1 and C2 co-referring to *F*. Further, for any *F**, it will be true that if *F* = *F**, then *F** is a functional/relational/dispositional property. I don't think this should be controversial: if a property is identical to another property, then their properties (which will be second-order now) have to be the same. This is a statement that is much weaker than even Leibniz's Law, which is its converse. Now, if Loar and others say that C1 refers directly to *F*, they have to accept that C1 is a physical-functional/relational/dispositional concept, which contradicts our shared view that C1 is not such a concept, but a monadic and categorical one. This is so for the following reason. A posteriori materialists say that phenomenal properties are identical to physical properties. Then we can take a phenomenal property *G*, and say that it is identical to *F*. Since, as I have pointed out, this means that *G* and *F* will have all their properties in common, we infer that the following is true about these properties:

> (G's property of) *being directly picked out by C1* = (F's property of) *being directly picked out by C2*.

Since a property being directly picked out by a concept makes the concept, by the meaning of 'directly', to be of the same kind as the kind of the property it picks out, we have the contradiction[10]:

> C1 is a physical-functional/relational/dispositional concept & C1 is a phenomenal/monadic/categorical concept.

[10] Note that I'm not denying the concept/property distinction, something a posteriori materialists insist upon, and think it's the key to understanding their theory.

And this completes my refutation of Loar's theory. In fact it points to the following dilemma for the materialist, if she wants to be consistent:

First Horn: Be qualia-nihilist = deny that there are phenomenal properties, and so you don't have to commit yourself to the existence of G, therefore you take C1 as an empty concept.

Second Horn: Be dualist = accept that what my argument shows is that phenomenal properties are not identical to physical/functional properties.

This is I think enough to show that the problems with Loar's account are quite serious. I now turn to the most sophisticated thinker in this area Stephen Yablo (1999). Yablo's defense of a posteriori materialism against the zombie argument turns on a fine-tuned distinction he puts forward, that between:

(α) It is logically possible that _____
(β) There is a logically possible world where _____

He thinks first that there is really a distinction here, and second, that zombies are to be put under the heading of (α) and not under that of (β). That is, as I have already said, he denies premise 4.: If zombies are conceivable, then zombies are logically possible. Where conceivability is the same as (α) and logical possibility is the same as (β). I have doubts as regards whether there is a real distinction here, and I suspect that it is mere playing with words. But I will try to show in a more systematic manner that Yablo's distinction presupposes a

primitive and unargued for notion of metaphysical modality and that his approach to modality is not sufficiently justified.

Yablo's understanding of the first statement is the same as what he calls "conceptual possibility", while I suppose the second statement could not be taken as otherwise than something having to do with what Yablo calls "metaphysical possibility". So let us make the two statements more explicit. That there is a logically possible world where there are zombies we could represent as follows (where P is the conjunction of all physical truths and Q is a phenomenal truth, such as 'There is consciousness', and where we use 'M' for the possibility modal operator):

(β *) There exists [world W_{LP} such that there are zombies, i.e. P & ~Q].

That it is logically possible that there be zombies we represent as:

(α*) M_{LP} [that there be zombies, i.e. P & ~Q].

Now that-clauses can be understood, as usual, as referring to possible situations (worlds)[11], so I think it is clear that with this interpretation the distinction proposed by Yablo collapses, because (α*) becomes:
(α**) M_{LP} [there exists [world W_{LP} such that there are zombies, i.e. P & ~Q]],

and I assume that logically possible existence of a logically possible world is the same as existence of a logically possible world, unless one wants to

[11] You don't have to be anything like a realist about possible worlds. The issue of Realism/Anti-Realism with respect to possible worlds is independent from the issues discussed in our context.

overcomplicate the modal space and say that there can be the case that it is logically possible that there is a logically possible world, but that does not entail that it is metaphysically possible that there is a logically possible world. If one appeals to such an idea, then she would have to explain why, and should put forth a reasonable motivation to double the space of worlds in this fashion.

Now Yablo may respond that that-clauses, or at least some of them, do not point to logically possible worlds (=metaphysically possible worlds), but to what we may call 'conceived worlds'. Let us see then whether this response is of any use. According to the proposal (α*) should be interpreted as and become:

(α***) M_{LP} [there exists [world W_C such that there are zombies, i.e. P & ~Q]]

In this formula a conceived world, W_C, should be understood as not committing us to the existence of a logically possible world. But now let us compare the following two formulae:

(1) M_{LP} $\exists W_C$: P & ~Q
(2) M_{LP} $\exists W_{LP}$: P & ~Q

Now if one considers that (2) will entail that $\exists W_{LP}$: P & ~Q, but (1) will not entail $\exists W_C$: P & ~Q, then the whole issue does not turn on the distinction between (α) and (β), as both (1) and (2) are prefixed by the same possibility operator. So, again, there will be reference to some primitive notion of metaphysical possibility that will do all the work. On the other hand, if both the above mentioned entailments hold, then our possibility operator is one that sanctions the inference to the existence of the entities it prefixes, worlds or

situations. So if this is so, then I don't see why M_{LP} (P & ~Q) does not just as well imply the existence of a zombie situation, or world, as it sanctions the inference to the truth P and ~Q holding at the same time. So, to reiterate my point from the start of this brief criticism of Yablo's position, it seems to me that his approach is quite artificial, and not (sufficiently) rationally justified.

4.3 CONCEIVABILITY AND THE EPISTEMIC/ONTIC DIVIDE: TWO-DIMENSIONAL SEMANTICS.

This section is dedicated to a more thorough analysis of epistemic and metaphysical modality. We saw that none of the epistemic arguments we discussed so far can successfully answered by a posteriori materialism. I will offer an a priori refutation of the zombie argument in the next section, but here I want to explain why a posteriori materialism fails in the case of the zombie argument as well. That is, I will explain why the appeal to a posteriori necessities is of no help for the materialist.

I already explained in Chapter II the idea behind Kripke's examples of a posteriori necessity, so I will not rehearse it here, but rather will expose what in my opinion is the most natural and rational interpretation of Kripke's idea, the two-dimensional semantic framework, proposed during the past 30 years in various sub-variants by authors like David Kaplan (1978; 1989), Robert Stalnaker (1978), Gareth Evans (1979), Martin Davies & Lloyd Humberstone (1981), Chalmers (1996, 2004), and Jackson (1994, 1998). I will rely heavily in this part on Chalmers (2004).

As I mentioned in Chapter II, Frege came to the conclusion, after exposing some puzzles regarding cognitively significant identity statements, that extension cannot be the whole story for an adequate semantics, that is, a theory of meaning, and postulated the existence of sense, *Sinn*. The basic idea was to delineate cognitively significant from trivial instances of identity statements, and the notion of sense served just that purpose. We can put this thesis in the following terms:

> ***Fregean Thesis***: Two expressions 'A' and 'B' have the same sense iff 'A = B' is cognitively insignificant.

By a cognitively insignificant claim we mean a claim that can be known trivially by a rational being. As a way to elucidate, that is, make more precise these notions, like cognitive significance, sense, and meaning, philosophers in the middle of the last century, most notably Rudolf Carnap (1947), proposed an analysis of them in terms of modal notions -- possibility and necessity. The notion of sense became that of intension, and that of reference became that of extension. Then, a thesis was needed to link the notions of intension and the modal notions, such as to give a criterion for the sameness of meaning. This thesis may be formulated as follows:

> ***Carnapian Thesis***: 'A' and 'B' have the same intension iff 'A = B' is necessary.

For an explication of Carnap's notion of intension it s worth quoting from Chalmers (2004):

> “Carnap's characterization suggests a natural definition: an intension is a function from possibilities to extensions. The possibilities here correspond to different possible states of the world. Relative to any possibility, an expression has an extension: for example, a sentence (e.g. 'All renates are cordates') can be true or false relative to a possibility, and a singular term (e.g. 'the teacher of Aristotle') picks out an individual relative to a possibility. An expression's intension is the function that maps a possibility to the expression's extension relative to that possibility. When two expressions are necessarily co-extensive, they will pick out the same extension relative to all possibilities, so they will have the same intension. When two expressions are not necessarily co-extensive, they will not pick out the same extension relative to all possibilities, so they will have different intensions. So intensions behave just as Carnap suggests they should.”

There is a structural isomorphism between intensions and senses, as intensions behave exactly as senses are supposed to. The difference is that intensions are tied to modality while senses to meaning. Intensions, understood in this way, provide a bridge between meaning and modality. Further, one can state one more condition for having a relation with reason as well. Kant provided such a condition or criterion, by way of the use of apriority for defining necessity:

> ***Kantian Thesis***: A sentence S is necessary iff S is a priori.

If we combine the Carnapian Thesis with the Kantian Thesis, we obtain the following:

Neo-Fregean Thesis: Two expressions 'A' and 'B' have the same intension iff 'A = B' is a priori.

This thesis captures the idea of what Chalmers calls “the golden triangle”, a link among the notions of meaning, modality, and reason:

> “The central connection between meaning, reason, and modality is captured within the Neo-Fregean thesis: intension is a notion of meaning, defined in terms of modality, that is constitutively connected to reason.”

Kripke’s idea of a posteriori came to break this triangle, more precisely, it severed the Kantian link between modality and reason, and by that the Neo-Fregean link as well.

Two-dimensional semantics, in turn, is a way to relink these three components: meaning, modality, and reason. It does so by offering a more refined approach to modality, which yields a Fregean aspect of meaning. According to the 2D semantics, there are two ways in which expressions pick out their extensions relative to possible situations. The first way of picking out extensions is via actual intensions of the expressions, that is, intensions that take a possible world as being the actual world, while the second way first acknowledges that the actual world is fixed, then considers the possible world as

counterfactual. This is why these intensions may be called actual and counterfactual intensions, respectively. As a result of taking into account these two ways to access the space of possible worlds there will be cases when an expression's evaluation with respect to a possible situation will give us different results, depending on which intension we use in doing so. Let us take an example for illustration: the sentence 'water = H_2O'.

We can test whether the negation of this sentence yields a possibility or not. We have two ways to approach the space of possible worlds, corresponding to the actual and conterfactual intensions of the terms involved, respectively. We will call these intensions 1-intension and 2-intension, respectively. A 1-intension for a term is more or less the same as a flaccid description associated with the term and is obtained when the possible scenarios involving its reference are considered as actual; it is the a priori aspect of its meaning. In our case 'water' has a 1-intension that may be captured by a long description like 'the transparent, colorless, odorless, drinkable, thirst-quenching, liquid that falls as rain, is to be found in lakes, rivers, seas, oceans…". The 2-intension of a term is more or less the same as the actual reference of the term, when the term is taken as rigid and the possible scenarios involving its reference are taken as counterfactual, and it represents the a posteriori aspect of the term's meaning. 'Water' then has a 2-intension which picks out its actual reference in every counterfactual world, that is, H_2O, as that is what science actually found out regarding what water is constituted by. The term H_2O, I take it, has coinciding 1- and 2-intensions, as it is a scientific term. Then, turning back to our sentence, we can observe that its negation is 1-possible, but not 2-possible. It is possible that "the transparent, colorless, …" is not H_2O, as a world where this description picks out some other chemical compound is conceivable (see the Twin Earth thought-experiment). On

the other hand, our sentence's negation is not 2-possible, as it is equivalent to the sentence '$H_2O \neq H_2O$', which is a contradiction. What is important if we take this semantics as capturing the ideas of Kripke and Putnam, and I think it is the best in capturing it, is that we should keep in view that there are always these two ways of analyzing a scenario when testing modal claims.

Let us now see the relevance of this admittedly brief discussion of 2D-ism for the issue of qualia and a posteriori materialism. A posteriori materialists usually insist on the Kripke cases of necessity a posteriori as cases where there is a broken conceivability-possibility link, and there is a strong rationale for separating the epistemic from the ontic perspective. Now 2D-ism can acknowledge this gap, but it gives it a more sophisticated and more adequate interpretation. These are case when there is a broken link, but not between conceivability and possibility as such, but between 1-conceivability and 2-possibility. At the same time, Kripkean examples do not show that there should be a similar gap between 1-conceivability and 1-possibility. As regards the epistemic/ontic divide, 2D-ism has a nice way to account for it. The epistemic is linked to 1-intensions, that is, to the a priori aspect of meaning, while the ontic to the 2-intensions, that is, to the a posteriori aspect[12]. But this does not mean that the epistemic fails to get hooked to possible worlds. There will be possible worlds corresponding to the epistemic aspect of the specific statements. In other words, the space of possible worlds will not be smaller as a result of the cases of a posteriori necessity, but rather it will be the same, but we will have two ways to access it. Hence, for instance, the scenario where water is not H_2O, will correspond to a bona fide possible world, namely, a world where the transparent liquid is not H_2O. We can see that the fact that the possibility of water not being

H_2O is epistemic, does not affect the status of the world where the 1-intension of such a statement is true: it is a perfectly good possible world, just as that in which water is H_2O. The zombie argument against a posteriori materialism that this semantic framework yields has been in recent years insistent upon by Chalmers, and can be put as follows[13]:

1. It is 1-conceivable that there be zombies. (epistemic premise)
2. 1-conceiving entails 1-possibility. (premise)
3. It is 1-possible that there be zombies. (from 1 and 2)
4. Physical terms have either coinciding or distinct 1- and 2-intensions. (analytic premise)
5. Phenomenal terms have coinciding 1- and 2-intensions. (premise)
6. If physical terms have coinciding 1- and 2-intensions, then zombies are 2-possible. (from 3, and 5)
7. If physical terms have distinct 1- and 2-intensions, then if zombies are not 2-possible, then the phenomenal truths are entailed by a deeper reality than that that revealed by physics. (premise).
8. Either zombies are 2-possible or the phenomenal truths are entailed by a deeper reality than that that revealed by physics. (from 4, 6, and 7)
9. If zombies are 2-possible, then materialism is false. (by the definition of materialism)
10. If the phenomenal truths are entailed by a deeper reality than that revealed by physics, then materialism is false. (by the definition of materialism)

[12] This is a bit simplified. In fact, this doesn't mean that the scenarios captured by 1-conceivings are not to be included in the category of the ontic. See below.
[13] This is my own formulation of the argument. It is I think more explicit than Chalmers's and by that better in capturing the possibilities regarding the mental/physical nexus.

11. Materialism is false. (from 8, 9, and 10)

The argument is valid. Let us see what premises one may doubt. The a posteriori materialist is in trouble, because she does not have any premise in this argument that would assert an entailment from conceivability to possibility *simpliciter*, but one that asserts an entailment from 1-conceivability to 1-possibility, premise 2, and that, as far as I see, neither Kripke, nor anyone having in mind the standard Kripkean examples contested. So if it is to be contested, it has to be based on an appeal to some non-standard a posteriori necessity. We will shortly and briefly discuss such an example, provided by Yablo (1999), but first let's continue with the analysis of the premises.

Denying premise 4 is equivalent to stating that there is a deeper reality than that revealed by physics. If besides denying this premise one also denies premise 10, then what she proposes is a nonstandard understanding of materialism. Daniel Stoljar (2001) proposes this line of thought. Denying premise 5 is the same as denying Kripke's insight that in the case of phenomenal concpets, like pain, one cannot distinguish the feeling of it from itself, as one can do in the case of non-phenomenal commonsense concepts. I think that insight is much more plausible than anything that could be brought in defense of its denial. Finally, one may deny the first premise. This is the line followed by a priori materialists, like Lewis, Dennett, or Shoemaker. This I will myself follow in the next section.

Let us now turn back to a posteriori materialism. As I have said, the denial of the entailment from 1-conceivability to 1-possibility is hard to justify. It would result in what Chalmers called a „strong necessity": a proposition whose negation is 1-conceivable, but which is nevertheless necessarily true. This

is why an a posteriori materialist will have to either postulate the lack of such entailment only in the mental/physical case or come up with some other examples in order to have an independent argument for strong necessities. Yablo (1999) considers the example of the equiconceivable existence and non-existence of a necessary God as one that will justify the belief in strong necessities. The concept of God is one that involves a necessary being. But we can equally conceive of his non-existence. Yet, if he exists, he exists necessarily. So it seems that we have a case of strong necessity.

As against this, we can argue, following Chalmers (1999), that here we have to use double modality, namely, conceiving something that is necessary, which would require the conceiver to reason meta-modally, in fact. If one sticks to the standard and most natural understanding of modal logic, the system S5, according to which iterated applications of modal operators do not affect the truth-value of the formulae, but they recollapse them to one operator prefixed formulae, Yablo's example is one when the conceiver should think beyond the level of the modal system, at a meta-level, which is not only unjustified, but inapplicable. It is one thing to imagine a God *simpliciter* and quite another to imagine a necessary God. As Chalmers rightly emphasises, most of us had (and some of us still have) doubts about the coherence of the very concept of God, even from an early age, and that is why we can imagine such a being as not existing, which means just that we can't imagine him existing necessarily, otherwise we would be required to go against the rational foundations of our modal claims. I will offer a case of strog necessity in the next, last chapter of this work, which is not controversial as Yablo's is. That I wil show to be compatible with our rational approach to modality, and the whole 2D system.

But you will have to wait a bit for that, and see what my response to the zombie argument is, which is just the topic of the next section.

4.3 THE INCONCEIVABILITY OF ZOMBIES.

As I have said, I want to offer a rationale for denying the first premise of Chalmers's argument. To understand my attack on it, we have to discuss a bit more thoroughly the notion of conceivability. I will rely on Chalmers 2002.

The most important distinction we have to consider is that between *prima facie conceivability* and *ideal conceivability*. A situation (or sentence) S is prima facie conceivable for a subject if, after reflection, there is no contradiction detectable in the hypothesis that S. Ideal conceivability is similar to prima facie conceivability, with the difference that we require ideal contradiction-proof rational reflection as its criterion. That is, we say that S is ideally conceivable if the ideal rational reflection by the subject does not reveal any contradiction in the hypothesis that S.

The notions of 1- and 2-conceivability we have already elucidated a few paragraphs above. Another distinction we should make is that between positive and negative conceivability. We will say that S is negatively conceivable if S is not ruled out a priori, or when there is no apparent contradiction in the situation or hypothesis expressed by S. Positive conceivability adds the condition that one be able to actually modally imagine the situation expressed by S. So positive conceivability is a more demanding kind of conceivability.

A special subcategory that is important to mention for my purposes is what Chalmers calls „secunda facie conceivability”. Let us quote him at this point:

> „A slightly better example of prima facie without ideal positive conceivability may be the Grim Reaper paradox (Benardete 1964; Hawthorne 2000). There are countably many grim reapers, one for every positive integer. Grim reaper 1 is disposed to kill you with a scythe at 1pm, if and only if you are still alive then (otherwise his scythe remains immobile throughout), taking 30 minutes about it. Grim reaper 2 is disposed to kill you with a scythe at 12:30 pm, if and only if you are still alive then, taking 15 minutes about it. Grim reaper 3 is disposed to kill you with a scythe at 12:15 pm, and so on. You are still alive just before 12pm, you can only die through the motion of a grim reaper's scythe, and once dead you stay dead. On the face of it, this situation seems conceivable — each reaper seems conceivable individually and intrinsically, and it seems reasonable to combine distinct individuals with distinct intrinsic properties into one situation. But a little reflection reveals that the situation as described is contradictory. I cannot survive to any moment past 12pm (a grim reaper would get me first), but I cannot be killed (for grim reaper n to kill me, I must have survived grim reaper n+1, which is impossible). So the description D of the situation is prima facie positively conceivable but not ideally positively conceivable.” […] [T]he Grim Reaper and impossible object cases are cases in a situation has not been coherently

> imagined. Of course in both these cases, the problem is revealed by a little reflection. One might say that in this case […], even if we have prima facie positive conceivability, we do not have secunda facie positive conceivability.

So secunda facie inconceivability, as I understood Chalmers, is something between prima facie conceivability and ideal inconceivability. It is a case when one apparently coherently imagines a situation S, but a little more reflection shows that, in fact, there was an important detail that the apparent conceiving left out, and so it is not in fact conceived that S. But we should also note that secunda facie conceivability is quite independent from ideal conceivability, which rerquires ideal reflection. I think it is a special category.

With this brief introduction of these notions, we can now turn to the analysis of the zombie argument. I think that I can make a good case, contrary to what all the authors involved in this debate held so far[14], for the prima facie positive 1-conceivability, but not secunda facie positive 1-conceivability of zombies. I want to avoid the whole idea of ideal conceivability, as I think it is a red hering in this debate, and to concentrate on prima facie and secunda facie cases of conceiving zombies.

[14] What has been held so far by philosophers involved in this debate are: (a) zombies are not even prima facie 1-conceivable (Dennett 1994, Lewis 1994, Shoemaker 19??), (b) zombies are prima facie 1-conceivable, but not ideally conceivable (Nagel 1974, 19??, McGinn 1999), (c) zombies are prima facie 1-conceivable, but not (2-)possible (Block & Stalnaker 1999, Balog 1999, Byrne 1999, Yablo 1999, 2000, 2002, Papineau 1998, 2002, Tye 1998, 2003), (d) zombies are prima facie 1-conceivable, but not prima facie 2-conceivable (or not ideally 1-conceivable, depending on the interpreatation) (Stoljar 2001a, b). The position I'm going to argue for shortly has not been put forth so far.

I think a potential gap between prima facie and secunda facie conceivability is usually identified, in philosophy, when some prima facie coherent scenarios are thought through, such that all the commitments of the conceiver are revealed, and then, finally, one of these commitments is contradictory. A case is that which I quoted from Chalmers, the Grim Reaper Paradox. A more controversial case is the conceivability of travel back in time. For instance, there are philosophers who think that this prima facie conceivable scenario commits the conceiver to the possibility of killing himself as a child, which in turn generates a contradiction, namely that one is both dead by the age of, say, 5, and that he is not dead at that age, since he is just conceiving of the scenario at the age of, say, 34. I think the zombie scenario is such a case, though what I will propose will surely be considered as controversial by those who are impressed with the zombie argument. In any case, I will try to do my best to convince them to the contrary. Strangely enough I will consider a thought experiment by an a posteriori materialist, Katalin Balog (1999), of which she thinks it shows the truth of a posteriori materialism, but I myself think it may be further developed to show the incoherence of the idea of zombies.

Balog imagines a zombie world where zombie Frank Jackson (the zombie counterpart of our Frank Jackson) exposes the 2D argument against a posteriori materialism, and tries to extract a contradiction from it as regards the conceivability-possibility link. Balog's strategy can be formulated as proceeding in three steps.

***Balog's strategy*:**

Step 1: Imagine the argument uttered simultaneously in the zombie scenario.

	Jackson (actual world):	**Zombie-Jackson (zombie-**

		world):
Premise 1	I am conscious	I am conscious+
Premise 2	If physicalism is true, for any truth T, there is a truth K, (expressing all the physical facts) such that 'K ⊃ T' is a priori (C-P thesis)	If physicalism is true, for any truth T, there is a truth K, (expressing all the physical facts) such that 'K ⊃ T' is a priori
Premise 3	'I am conscious' is not a priori entailed by K (i.e. it is conceivable that K & not 'I am conscious')	'I am conscious+' is not a priori entailed by K (i.e. it is conceivable that K & not 'I am conscious+')
Conclusion	Physicalism is false.	Physicalism is false.

So the basic idea is that Jackson's zombie counterpart, Zombie-Jackson, will expose the same argument, using the same words, during his talk at Zombie-Oxford University, with the only difference that his phenomenal terms corresponding to the Jackson's phenomenal terms will have to refer to something else, as there are no phenomenal properties instantiated in the zombie world.

Step 2: Analysis of logical commitments.

But the conclusion is false in the zombie-world, since physicalism is true in that world, by stipulation. Therefore, Zombie-Jackson's argument is not sound, so one of the zombie premises has to be false. Accordingly, one of Jackson's premises has to be false, on the plausible assumption that the zombie has intentional states and so what he says is not meaningless.

Step 3: Argument for the falsity of the a priori entailment thesis and conclusion regarding physicalism.

Balog's conclusion: the false premise is premise 2. More to the point, phenomenal concepts are special in some ways (see Loar above) that make the conceivability-possibility link questionable when they are involved. So it hasn't been shown that physicalism is false.

Chalmers' reply

The best candidate for being false is premise 1, 'I am conscious' as uttered by the zombie (but not, of course, by Jackson). As he puts it[15]:

> This is an intriguing argument, but I think the problem with it is clear. Balog's parallel argument requires that a zombie's claim "I am conscious" is true; otherwise the argument doesn't get off the ground. Balog supports this by suggesting that the zombie's "consciousness" concept will pick out a physical/functional property to which it is causally related. But I think it is much more plausible that the zombie's claim is false. The easiest way to see this is to consider an argument in the zombie world, perhaps between Zombie Chalmers and Zombie Dennett. Zombie Chalmers says "Qualia exist", Zombie Dennett says "Qualia do not exist". Balog's analysis implies that in the zombie world, Zombie Chalmers is right. But this seems wrong. Surely in the zombie world, at least, Zombie Dennett is right.

[15] This is Chalmers's' unofficial response to Balog, to be found at: http://www.u.arizona.edu/~chalmers/responses.html#balog.

Well, I think Chalmers is perfectly right as regards the most plausible prima facie candidate to be denied in Zombie Jackson's argument. But it is not I think the best secunda facie candidate. My strategy for attacking the zombie argument consists of four steps. And my main point is to draw a quite different lesson from Balog's argument, one that sanctions the denial of the third premise in Jackson's and Zombie Jackson's arguments, or that of the first premise in my reconstruction of Chalmers's zombie argument, or the third premise in my initial general formulation of the zombie argument.

My strategy:

Step 1: Acceptance of C-P principle.

I accept that *the right kind* of conceivability entails *the right kind* of possibility. That is, ideal positive 1-conceivability entails 1-possibility (Chalmers 2002). This means that I don't want to contest, as a posteriori materialists do, the conceivability-possibility link.

Step 2: Clarify what is involved in the case of zombie utterances in terms of concepts.

Follow Chalmers' own model of taking a world as actual when evaluating 1-possibility. Then, we have to take the zombie-world as actual. But since the zombie's utterance of 'I am conscious' involves the zombie world's *mental reality*, unlike his other utterances, like, say, 'There is water', and since

Chalmers himself (see his criticism of Stalnaker's contextual interpretation of 1-intension, 2002) refuses to endorse a metasemantic understanding of 1-intensions, but adopts a semantic view (for instance, he does not say that if the a watery stuff contains no oxygen, we call it "water"', but that if the watery stuff contains no oxygen, then water contains no oxygen[16]), we have to accept that when evaluating the truth-value of the zombie's utterance what is relevant is the zombies *concepts*, which I take to be *conceptions* or *ways of thinking* about a referent[17].

Step 3: Zombies are not irrational but their thinking conforms to some rules of rationality.

There are I think some rules of rationality clearly and non-controversially applicable to zombies[18]:

a. *Non-contradiction*: they do not assert any explicit contradiction (p & not p)
b. *Implication*: They know that if (p & not q), then not (p ⊃ q).
c. *Relevance*: if they know that p ⊃ q, and they assert p, they will conclude q, without first asserting another proposition r.

Step 4: Argument for denying the premise that phenomenal truths are not entailed by the conjunction of all physical truths.

[16] The indicative conditional reflects the fact that we evaluate 1-intension at a world, that is, we consider it as actual.
[17] On a metasemantic view, on the other hand, we would consider the zombies term as *mentioned*.

Consider Zombie-Jackson's argument. Chalmers says that the first premise has to be false; this is ok, but let us see what the consequences are.

Suppose, as Chalmers does, that 'I am concsious+'

a.) **is false.**

This can be so only if one of the following two hypotheses is true:

a1.) 'conscious+' expresses the same concept as our concept *conscious*, in which case the zombie is conscious: contradiction.

a2.) 'conscious+' expresses the concept 'a fact/property over and above the physical facts/properties of the actual world', in which case dualism follows without any need for the next, second premise, so the zombie's behaviour will diverge from ours, because he will not assert the next premise, since it would be irrational to do so (by the rules of rationality we have established): contradiction.

Therefore, 'I am concsious+'

b.) **has to be true.**

But this means that Balog's argument is OK, from which I, contrary to Balog, further infer the following argument (where Z means 'there be zombies'):

[18] See also Sidney Shoemaker 1999.

1. It is conceivable that Z only if conceivable that Z does not entail possible that Z. (from the conclusion of Balog's argument).

2. Either (Conceivable that Z entails possible that Z), or not. (analytic premise)
Hence,
3. Either it is not conceivable that Z, or conceivable that Z entails possible that Z.
4. Conceivable that Z entails possible that Z (instantiation of the C-P principle; it would be implausible for it to hold in all cases, but not in this one, as Chalmers (1996, 1999, 2002) and myself in this chapter point out).
Hence,
5. It is not conceivable that Z.

Then given all this:

1. 'I am conscious+' is either true or false. (analytic premise)
2. If true, then zombies are inconceivable. (by the argument devised at point b.) above)
3. If false, then zombies are inconceivable. (by the contradiction established at point a.) above)
4. Zombies are inconceivable. (from 1., 2., 3.)

This means that it is neither the first (Chalmers), nor the second (Balog) premise that is false in the zombie world, but the third. More importantly, there are independent reasons to consider that premise 3 is the best candidate to be false.

Namely, first, as against the Balog interpretation, when considering whether something is possible, we first try to conceive of it. That is, we put our concepts to work But not all of our concepts (conceptions) are easily and univocally superficially identifiable. When trying to conceive of something, we simultaneously test our superficial understanding of some concepts. This is why we devise thought experiments. This is an exercise in digging deeper into our real conceptual commitments and their inferential liaisons. I take Balog's strategy as being such an exercise, meant to test whether zombies are conceivable (unlike Balog herself, who takes their conceivability for granted). Now it is implausible, if not odd, to say, after taking the steps along Balog's line of thought, that we have conceived of zombies, but found that something has to be wrong with conceivability of zombies argument, and further, that conceivability does not lead to possibility. It is more natural to say that we have not succeeded to complete the zombie scenario, that is we have not been able to conceive them. There was a prima facie conceivability of them, but after taking some further steps, we have realized that we were wrong. Premise one is the best *prima facie* candidate, but following the train of thought I have exposed, premise 3 is the *secunda facie* candidate.

Second, as against the Chalmers interpretation -which he puts in a nutshell as 'In the debate between Zombie-Chalmers and Zombie-Dennett, the latter is right'- we could consider a further development of Balog's story. Consider that in the Zombie-Jackson world there are also Zombie-Balog and Zombie-Chalmers. Zombie-Balog will expound the same argument by considering a Zombie-Jackson* (a zombie with respect to their world), and Zombie-Chalmers will say that 'I am conscious++' has to be taken as false. Now whatever Zombie-Jackson's or Zombie-Jackson*'s concept of consciousness, it

is clear that it is hard to find a non-arbitrary reason to consider that Zombie-Chalmers is right as against Zombie-Balog regarding the truth-value of Zombie-Jackson*'s utterance of the first premise. One possibility that I think is interesting is to say, roughly, that we decide the truth-value as a function of the role the consciousness-concept of various utterers play in the relation between their actual world and their conceived scenarios. Chalmers himself at some point in his book (1996, ??) talks about the possibility of a somebody's having a concsiousness-concept that fulfills the role of our concept, but does not have the same content. I think this represents a possibility worth taking seriously, but it leads to a conceptual-role understanding of the concept of consciousness itself.

As a consequence, we can accommodate both what is plausible in Balog's and in Chalmers' interpretation: zombies are prima facie conceivable (concession made to Chalmers), but prima facie conceivability does not entail 1-possibility (concession to Balog); finally, zombies are not secunda facie conceivable, so they are not possible. Balog's otherwise very nice idea of a thought experiment yields an argument against the soundness of the zombie argument, but not for the reasons she thinks it does, but rather those I have tried here to expose.

There remains one thing to do, as I did at the end of each chapter, namely, to formulate a further worry, one that, together with the worries expressed at the end of the previous chapters, will receive a rational answer in the last chapter, when I expose a new theory of the mental/physical relation.

Problem 3: Why do zombies seem to be conceivable? Why do we have the prima facie intuition that zombies are possible?