Contents

Knowledge, Possibility and Consciousness

John Perry
Stanford University

August 2, 1999

Chapter 1

The Subjective Character of Experience

(This document contains chapters 4, 7 and 8 of a book I am working on. I've left the other chapter headings in so the reader can get some idea of the rest of the book from the table of contents. The full text of the current draft can be downloaded at http://www-csli.stanford.edu/ john/NICOD/nicod.html. Comments can be sent to john@csli.stanford.edu and will be much appreciated. –John Perry

Chapter 2

Sentience and Thought

Chapter 3

Thoughts About Sensations

Chapter 4

The Zombie Argument

4.1 The Possibility of Zombies

As the first step in his Zombie Argument, David Chalmers invites us to consider what he describes as a logical possibility:

> ...consider the logical possibility of a *zombie*: someone or something physically identical to me (or to any other conscious being), but lacking conscious experiences altogether. At the global level, we can consider the logical possibility of a *zombie world*: a world physically identical to ours, but in which there are no conscious experiences at all. In such a world, everybody is a zombie.
>
> So let us consider my zombie twin. This creature is molecule for molecule identical to me, and identical in all the low-level properties postulated by a completed physics, but he lacks conscious experience entirely. (Some might prefer to call a zombie "it," but I use the personal pronoun; I have grown quite fond of my zombie twin.) To fix ideas, we can imagine that right now I am gazing out the window, experiencing some nice green

> sensations from seeing the trees outside, having pleasant taste experiences through munching on a chocolate bar, and feeling a dull aching sensation in my right shoulder.
>
> What is going on in my zombie twin? He is physically identical to me, and we may as well suppose that he is embedded in an identical environment. He will certainly be identical to me *functionally*: he will be processing internal configurations being modified appropriately and with indistinguishable behavior resulting. He will be *psychologically* identical to me...He will be perceiving the trees outside, in the functional sense, and tasting the chocolate, in the psychological sense. All of this follows logically from the fact that he is physically identical to me, by virtue of the functional analyses of psychological notions...It is just that none of this functioning will be accompanied by any real conscious experience. There will be no phenomenal feel. there is nothing it is like to be a zombie.([Chalmers, 1996]: 94-95)

According to the Zombie argument, then, it is logically possible that there be a world in which people are exactly like us in every physical detail, but do not have experiences, or have experiences that are not like anything to have. These people would be indistinguishable from us in terms of behavior and physical structure down to the last detail.

4.2 Why zombies could not be physically like us

I'll use the term "Zombie world" for a possible world in which there is no consciousness, but there are creatures that look and act like us and are like

us inside, *insofar as this is possible given the lack of consciousness.* That is, a Zombie World will be just like ours except for the conscious states, *and* whatever other differences the lack of conscious states implies.

I'll use the term "Chalmers' Zombie World" for a world that is a zombie world, and is, as Chalmers' argument requires, is physically *indiscernible* from ours.

From the point of view of an antecedent physicalist, it seems that Zombie worlds are possible, but Chalmers' Zombie worlds are not. The reason is that the antecedent physicalist believes in the Efficacy of the Conscious, and rejects epiphenomenalism. Since the antecedent physicalist thinks that conscious mental states bring about changes in the world, it seems that a world without them will have to differ in some way from ours. Either the changes won't occur, or they will occur, but will be caused by something else. If conscious states make a difference in the way our bodies work and ultimately in how we behave, and they are absent in the Zombie world, then how could everything in the physical world be the same as it is in our world?

An analogy. We can imagine a world like ours, but with no water. But we cannot imagine a world with no water, and everything else the same. If there were no water, there would be no plant growth, no floods, and so forth and so forth. We might imagine a world just like our world was on July 1st, 1955, with all of the water suddenly or gradually removed. For some reason, let's suppose, the process of condensation ceases, although evaporation continues. As time passes, the lack of water in that world will cause it to diverge in more and more major ways from our world. The plants will die, the fish will die, the people will die, and so forth and so on. This would be true whether or not water was reducible to hydrogen and oxygen, or, contrary to fact, were a perfectly separate substance not

further reducible. Still, if water plays a causal role, and you remove the water, everything else will not be the same.

If we removed the conscious states from our world, say just as it is right this minute, as we imagined doing with the water, what would happen? We leave all of the (other) physical states intact, and all the of the laws of nature intact, except those that have conscious states as effects. What will this world be like? If we believe in the principle of the Efficacy of the Conscious; that is, if we believe in our world conscious states make a difference, then we will think that this Zombie world will begin to diverge from ours. Consider the case of me picking up the white-hot piece of charcoal. In the Zombie world I will not feel the pain, as I do in this one. So the things that that feeling of pain causes, such as memories of a certain sort, will either not occur, or will occur for different causes. In either case, our world will have to be different from the Zombie world.

Suppose I bite into a fresh, warm, chocolate chip cookie. I am in the state of being somewhat hungry, and, for some reason, not worried about my weight or other health matters. I taste the chocolate chip cookie, in the phenomenal as well as the psychological sense. I attend to the what-its-like property of my brain state—although of course it seems very much like something wonderful happening in my mouth. I say, "Boy, was that good!!". Now I find it simply incredible—not inconceivable, but really quite incredible—that the conscious event was not part of the cause of my saying what I did. It seems to me that if some other conscious event occurred, like the kind of conscious even that occurs when one chews on zinc-coated nails, I would not have said what I did at all. And it seems to me if no conscious events and ensued upon chewing the cookie—if my stream of consciousness had just continued on, with no new taste sensations—I would have been surprised and disappointed, and would not have said,

"Boy, was that good!!" So it seems to me that a cause of my remark, an INUS condition in Mackie's terms, an Insufficient but Necessary part of an Unnecessary but Sufficient condition, was the conscious event.

Now let's consider my Zombie twin. We are asked to suppose that my Zombie twin puts the cookie in his mouth, chews it up, and says, "Boy is this good!!" But what will make him say that, if there is no conscious state, no burst of chocolate chip cookie flavor in his mouth?

Of course, Zombie-John might utter the sentence "Boy was that good!" The same observable events might happen in the Zombie world as in the actual one. It might happen as a result of a different cause, or simply occur, with no cause at all. The antecedent physicalist can certainly suppose all of this to be logically possible, without in any way compromising the view that the conscious state is a physical state of the brain, for such a world will not be physically indiscernible from ours, and hence not a Chalmers Zombie World.

So we need to be careful of the difference between simply imagining a Zombie world and imagining a Chalmers Zombie World. Consider any specific event that we suppose is caused in part by a specific conscious state. Call the event X. Suppose X is caused by the combination of A,B, and C. A and B are the physical causes and C is a conscious state. Together they are a sufficient condition for the physical event X, and each is a necessary part of the sufficient condition. In the Zombie world C, the conscious event doesn't occur. So if the Zombie world works just as ours does, X won't occur either, because the physical conditions, without C, are not sufficient. And so the Zombie world isn't just like ours. But of course we can imagine X occurring in the Zombie world, even though C doesn't occur. X can just occur. Why not? It could just occur sort of miraculously, or it could be that the physical principles of the Zombie world are different than the

actual world, so that A and B are causally sufficient for X. So again, the Zombie world isn't physically just like ours. In our world, X occurs, caused by the combination of A, B and C, and A and B alone are not physically sufficient for X.

I want to mention two possible misunderstandings. First, I am not claiming that we are always right about the effects of our conscious states. Suppose I perform Ewing's experiment, and pick up a piece of white hot charcoal. I feel pain, I drop the charcoal. It seems to me that the feeling of pain caused me to drop the charcoal. It may be that I am wrong about that. It may well be that I drop the charcoal, quite independently of the feeling of pain; that the feeling of pain, and the release of the muscles that hold the charcoal, are both caused by more immediate effects of the heat of the charcoal on my nervous system, rather than the pain being the cause of the release, as it seems. There is no reason for the Antecedent Physicalist to think that we are always right about what conscious states cause.

But note that in a case like this, the feeling of pain will have other effects. The next time someone suggests that I pick up a piece of charcoal, for example, I will be very reluctant, because I remember what the pain was like, and it is would be very hard to accept that the memory of what it was like, did not depend on what it was like, and that the influence of the memory, was not connected to the nature of the memory—to what it is like to vividly remember picking up the charcoal. It is very hard to accept that if the experience of picking up a piece of white hot charcoal was like the experience of eating a warm chocolate chip cookies, that I would not at least be tempted to perform the experiment again.

The second possible misunderstanding is this. It might seem that I am saying that a certain world isn't possible, for contingent reasons. That is, because antecedent physicalism happens to be true, a contingent fact,

the Chalmers Zombie World isn't possible. But what is possible or not shouldn't depend on contingent facts.

Part of the answer to this objection will depend on issues about identity, necessity and conceivability, which I'll consider in chapter 8. But the basic point is simply this. A Chalmers Zombie World is not simply a world in which various things occur. It is certainly possible that there be a world with all of the same events as ours, except for the conscious events. That is not enough for it to be a Chalmers Zombie World. The second condition a Chalmers Zombie World has to meet is being physically indiscernible from ours. That is a matter of having a certain similarity to the world that happens to be actual. Whether a given possible world qualifies as a Chalmers Zombie World, then, is not simply a matter of what happens in it, but also a matter of its similarity to the actual world. So whether a given possible world qualifies as a Chalmers Zombie World depends on contingent facts about the actual world, namely, what the actual world is like. The antecedent physicalist simply claims that none of the possible worlds both meet both of the conditions of being a Chalmers Zombie World. The point is not that if the causal facts are different, some world not logically possible that otherwise would be. The point is that if the facts about the causation of non-conscious events are different, then we are not imagining a Zombie world that satisfies Chalmers' directions, but some logically possible world in which both the conscious states and many of the other physical facts are different.

Now this is not too surprising. The antecedent physicalist supposes that the what-it-is-like properties *are* physical properties. So clearly the Antecedent Physicalist will find a problem in the claim that there is a logically possible world that is physically indiscernible from ours, but in which no one has any what-it-is-like states.

Table 4.1: Two Separate Issues

	Epiphenomenalism	**Efficacy of the Conscious**
Physicalism	Conscious states are physical nomological danglers, — in principle publically observable	Antecedent Physicalism
Dualism	Chalmers' Position: Conscious States as non-physical nomological danglers	Common Sense Dualism: the physical world world is not a closed system

4.3 Dualism and Epiphenomenalism

What may be somewhat surprising though, is that the possibility of a Chalmers Zombie World really has virtually nothing at all to do with the issue of physicalism versus dualism. The issue for which it is a test, is epiphenomenalism versus efficacy of the conscious. The two issues are independent. Table 3.1 shows the various possibilities.

Epiphenomenalism, the doctrine that conscious events are effects but not causes, is consistent with physicalism. And if one is a physicalist and an epiphenomenalist, one will accept the possibility of the Chalmers Zombie world. The world will not be physically indiscernible from ours, but it will be physically indiscernible from ours, except for the absence of consciousness, and that is the possibility Chalmers invites us to consider. Even for epiphenomenalists, the Zombie argument does not provide an argument for dualism.

On the other hand, one can be a dualist, and accept the Efficacy of the Conscious. Indeed, this may be common sense, and it has certainly

been philosophical common sense throughout certain periods of history. It is natural to believe in the efficacy of the conscious, and, because of the intuitions captured by the Leibniz and Ewing thought-experiments, dualism is natural too. There is nothing inconsistent about this position. Its advocate would find the Chalmers Zombie World quite impossible, for just the same reasons the Antecedent Physicalist does. Since conscious events make a physical difference, the physical world, without them, cannot be indiscernible from our own. The problem with this view is mainly that the arguments for it, however intuitive their force, are simply not compelling, and it denies that the physical world is a closed system, that is, that physical events have only physical causes.

All four boxes in Table 3.1, then, are occupied by logically consistent positions. My point has not been that Chalmers's view is impossible, but only that the Chalmers Zombie World is. Of course, *if* one is an epiphenomenalist, *then* it will not seem impossible that a world could be without conscious experiences and yet (otherwise) physically indiscernible from ours. But the acceptance of the possibility of such a Chalmers Zombie World still does not provide an argument for dualism, for it should be as acceptable to the physicalist epiphenomonalist as the dualist epiphenomenalist.

A Chalmers Zombie world, then, seems to be a test for dividing epiphenomenalists from non-epiphenomenalists, not an argument for defending dualism against physicalism. All epiphenomenalists pass the test of finding the Chalmers Zombie World conceivable, all non-epiphenomenalists fail it. Dualists and physicalists will both pass the test if they are epiphenomenalists, and fail it if they are not.

At most, then, the Zombie argument is an argument for epiphenomenalism. But it is not a very convincing one. If there is a Chalmers Zombie World, then epiphenomenalism must be true. To show that there is a pos-

sible world meeting certain conditions, one must imagine or describe it in enough detail to be sure it is possible, and meets the conditions in question. We can surely describe a Zombie World, but to meet the conditions to be a Chalmers Zombie World it has to be physically indiscernible from the actual world, except for the absence of conscious events. What reason would we have to suppose that among the possible worlds meeting the conditions of being Zombie worlds, there is one that meets the further condition of being a Chalmers Zombie World? I cannot see any reason we would think this, unless we were *already* epiphenomenalists.

4.4 Supervenience and Epiphenomenalism

I've oversimplified Chalmers so far, in an important way, by leaving out the topic of supervenience. If we go back to the quote with which I opened the chapter, we find that the conditions on the Zombie world seem to shift a bit from the first paragraph to the second. In the first he says the zombie world is "physically identical to ours," but in the second paragraph he says,

> ... let us consider my zombie twin. This creature is molecule for molecule identical to me, and identical in all the low-level properties postulated by a completed physics, but he lacks conscious experience entirely ...

So what is the Zombie world supposed to be like? Is it physically indiscernible? Or is it just indiscernible with respect to the low-level properties postulated by a completed physics?

Many physicalists assume that if world w_1 and world w_2 do not differ in the low level properties postulated by a completed physics, they will not differ in any of the higher level physical properties either. The higher level properties and the existence of the complex objects that have them all

have to do with the way the basic particles and their properties fit together. That is, once you've got all the events happening at the most basic level (which we usually think of as the smallest in size and shortest in duration) and all the basic relations between the basic things, you have all of the rest. For a non-basic physical fact to obtain, is just amounts to a certain complex combination of basic physical facts obtaining.

A theological metaphor borrowed from Kripke [Kripke, 1980/97] may be helpful here. By Thursday of the week of creation, God has decided just what all the molecules, or atoms, or quarks, or whatever the bottom level of stuff is, will be doing, where, and when. Now, does he have to come back the next day and decide if the Atlantic Ocean will be salty, or if there will be snow on Mount Everest? No, his work is done, as far as the physical part of the world goes.

Given that picture, there is no real difference between the requirements of the first paragraph and the requirements of the second. Why then the difference in formulation?

I think Chalmers wants the physicalist to focus on the question of where he can put the phenomenal properties. Will God's work up through Thursday determine when and where they occur? Or will he have to go back to work Friday and make those decisions? It seems that the phenomenal properties must be in one of the following categories:

A) Low-level properties postulated by a completed physics, which I'll call "basic physical properties".

B) Complex physical properties: properties that can be identified with conjunctions, disjunctions or other first-order logical constructions from basic properties. A) and B) together I'll call "First-order physical properties".

C) Second order physical properties; properties of the form "has a first-order physical property that meets condition C" where whether a property meets condition C depends only on first order physical facts. These properties "logically supervene" on physical facts. (The use of "logic" is a little confusing; it is used here is a somewhat broader and looser sense than in B). Logical supervenience is contrasted with causal supervenience; the latter calls for new facts, the former only new ways of organizing and classifying them, ways that may go beyond the strict techniques implied by logic in B. If one understands the principle of classification one can see that the supervening property is present in certain situations, simply as a matter of meaning or logic, broadly conceived.)

If the phenomenal properties are in any of these categories, God is done Thursday evening. He doesn't have to come back to work Friday to decide what to do about them. But he has more work to do if phenomenal properties belong in either of the following categories:

D) properties that are not in B or C, but causally supervene on A.

E) properties that neither logically nor causally supervene on A.

Let's eliminate E) as a possibility for phenomenal properties, as contrary to the overarching scientific hypotheses of our time (and at any rate not the position of either Chalmers or the antecedent physicalist). That leaves A) - D). A)-C) would leave subjective characters as clearly physical properties. C) differs from A) and B), however, in that the subjective characters could not be *identified* with physical properties, neither the basic ones, nor those definable from them by logical techniques. Still, if a brain states having a subjective character simply amounts to its having certain basic physical

properties, then, even if for some reason the exact combination required can't be captured by logical techniques, we don't seem to have a property that is non-physical in any respect that has much metaphysical bite. It might be, for example, that the property of being a valve belongs in class C), but the existence of valves still wouldn't seem to be very interesting from a metaphysical point of view.[1]

If property belongs to class A), B) or C), then, there will not be two logically possible worlds, indiscernible in terms of basic physical properties and the laws that govern them one of which has property and other other of which does not. If it is case C), a logically supervenient property, the occurrence of the supervenient property is not an extra fact or existence; having the physical goings on amounts to having the supervenient property.

D) requires something more than this. If a property is causally supervenient, there will be pairs of logically possible worlds, physically indiscernible at the level of basic physical properties and the laws that govern them, in one of which the property is exemplified, and in the other of which it is not. By late Thursday afternoon, God will have narrowed down the world he was going to create to a set of worlds, ones that are physically indistinguishable, alike in their A), B) and C) properties, but different in their D) properties. God will have to add a law or laws to nature, saying that in certain physical circumstances, these properties will occur. Their occurrence will be determined by the physical properties, but will not simply amount to the occurrence of the physical properties, but be something more, something

[1] One might argue that valves are pretty important, because one thing valves have in common is a certain role in the lives of beings with minds. Perhaps then there being valves implies the existence of things with minds. So, if minds are non-physical, valves imply dualism. But then we might as well think about minds directly, and not take a detour through valves. To return to the theological point of view, if minds are physical, then God didn't have to work on Sunday to decide which objects get to be valves. If minds are not physical, then as long as he spends Sunday deciding what minds are like, valves will be taken care of.

additional. D) properties seem to be a version of what used to be called "emergent properties".

The target of the Zombie argument, I think, is a philosopher for whom the live choices are C) and D). There are two reasons why it might seem fair to ignore A) and B). First, it seems that twin arguments and multiple realizability arguments have convinced most philosophers that A) and B) are not viable; the most clearly physical status the physicalist can plausibly claim for mental states is some kind of supervenience. Second, since supervenience is a weaker form of physicalism than identity, if we can eliminate C) as a possibility, we don't need to worry about A) and B).

Let's review the reasoning behind the move to supervenience. We'll start with a pretty plausible case, the property of being a valve.

Why do we suppose that a property like being a valve might be only logically supervenient on basic physical properties, rather than fully reducible to them? It seems that we usually have one or both of two things in mind. First, the question of whether something is a valve (or a dollar, or a husband, or a sentence of English) might not depend on just the local physical properties of the thing, but on various contextual and historical facts: how it was created, where, and the like. Twin arguments bring home this point. One might have two identical structures, one of which was a valve, and one of which was device for pitting prunes. The valve would be a valve in virtue of the reason for which it was made, who made it, where it was sold, and what it was used for, and the prune pitter would be a prune pitter for analogous reasons. You might be able to use the prune pitter for a valve, perhaps you could even turn the prune pitter into a valve. But the prune pitter pitting prunes is not a valve, even if its structure is identical to the valve in the next room controlling the flow of water into the washing machine.

The second point, is that because it is the capacity to perform a certain function that makes a thing eligible to be a valve, things with indefinitely many physical configurations and compositions might serve. Multiple realizability examples make this point. Two structures that are quite different might both be valves, because of they were manufactured, sold, bought and used to control the flow of water.

We have then cases of "physical twins", that differ in certain properties, which depend on historical and contextual factors. And we have dissimilar physical things, that share properties, because they can perform the same function. There are just lots of ways to be a valve. In such a case, it seems that a straightforward identification of the property of being a valve with a basic physical property or even a first-order physical property will likely not be possible.

It seems that many mental properties, the ones Chalmers calls "psychological properties", are like being a valve in both ways. Twin arguments point to the non-local, externalist nature of many such properties. (Recall the example involving Moravcsik and Flickinger in the last chapter.) Multiple realization cases point to the functional nature of many mental properties. In all of these cases, it seems that logical supervenience, level (C), is the appropriate relation between the physical world and the mental states. If we fix all of the physical facts, a physicalist will claim, we will fix these functional facts. So physicalism can be true, even if we identify the psychological states with basic for first-order physical properties.

Consider our standard philosophical Martian and me. Both of us can be in the *psychological* state of pain, even though our brain states are not the same. What we have in common is that the quite different states we are in, share some (suitably abstract) causal role. We both have a barrier between us and the outside world; mine is skin, his is something else. We both have

ways of exiting situations. We both have ways of getting help from others. And we both have an internal state, that typically occurs when our barrier is stressed, and typically leads to attempts to exit and/or get help. Our two quite different states share the causal role of pain. The psychological state of pain then logically supervenes on the first order physical properties. So far so good.

Suppose now that we were convinced of two things. First, that the Martian and I, since we were functionally just alike, not only were both in the *psychological state of pain*, but were also in the same *phenomenal* state. What it was like for the Martian, when he stepped on the tack, which almost punctured his barrier to the outside world, was just what it was like for me, when I stepped on the tack, that almost punctured my skin. Second, that, as was deemed common sense two chapters back, what it was like for me, and what it was like for the Martian, depends on what goes on inside us, at the moment of pain; that the what it is like aspect of the state is not a causal, historical, or functional property.

If we adopt C) with respect to subjective characters we can get the first thing we want. We can say that not only the psychological state of pain, but also the phenomenal state of pain, logically supervenes on causal role and function. So the Martian and I are in the same phenomenal state.

But this won't get us the second thing we want, that my experiences are a matter of what is going on inside of me, not a matter of how what is going on inside of me fits into the rest of the world.

The Martian and I are in different first-order states. We are in the same second-order causal/functional state, but that does not suffice to put us in the same phenomenal state, if that is a local, non-functional state. Something more is required. Using the theological metaphor, we require a decision by God to grace the Martian's Mars-brain states, and my human

brain states, with the same subjective character. God has to decide that functionally equivalent states should have the same functional character. But that would amount to D), causal supervenience. And of course if God could have made the decision to grace us both with the same subjective characters, it seems he could have made the choice of gracing neither of us within any qualia at all: the Chalmers Zombie World. So subjective characters are not identified with functional states, but causally supervene upon them.

If I were convinced that C) or D) were correct, and I had to be either a functionalist about subjective characters, contrary to common sense, or a dualist, I would either go for D), or take early retirement.

But I don't see any argument for the restriction to C) and D). The reasonable way out of this dilemma between C) and D) is to ignore it, and choose B). Subjective characters are first-order physical states. We should reject supervenience, and accept an identity theory for phenomenal states. We should reject both C) and D), and accept B).

This means that we will have to accept that the Martian and I are not in the same phenomenal state. But what reason is there to suppose that we are? It seems to me that whatever reason we thought we had, was based on ignoring the Block-Chalmers distinction between psychological states and phenomenal states. Can we accept this consequence?

To suppose that the Martian and I are not in the same phenomenal states, is not necessarily to deny that Martians have phenomenal states. Some of the internal states of Martians may be like something to be in. We may find that our psychology fits the Martian very well. We may find we can predict and control our Martian using the same basic framework of desires, intentions, emotions, beliefs, goals, fears and the like as we use for ourselves. If so, there will be a place in his psychology for pain and

pleasure, for our psychology could not begin to fit onto being that were not motivated by pleasures and pains.

One often compares Martians and robots in discussions of supervenience, as two sorts of alien beings, with respect to which to whom the denial or affirmation of consciousness might be an issue. But there are big differences. Martians would presumably be naturally occurring beings, evolved on Mars. If we find our belief and desire psychology fits them, we have reason to suppose that the basic architecture of their mentality is like ours; that their intentionality is, as Searle says, "natural" and not manufactured. With any robots that now exist, or are likely to, the case will be quite different. Their susceptibility to intentional description will have been planned by their creators. I do not mean to say that robots could not have natural intentionality, and could not have, what seems to me a requirement of it, phenomenal pains and pleasures that their basic architecture motivates them to avoid and seek. But I see no reason to suppose that the robots now envisaged do so.

4.5 The Inverted Spectrum

After his exposition of the Zombie argument, Chalmers notes that such a dramatic possibility is not required for the dualist argument.

> It suffices to establish the logical possibility of a world physically identical to ours in which the facts about conscious experience are merely *different* from the facts in our world, without conscious experience being absent entirely. As long as some positive fact about experience in our world does not hold in a physically identical world, then consciousness does not logically supervene.([Chalmers, 1996]: 99)

> It is therefore enough to note that one can coherently imagine a physically identical world in which conscious experiences are *inverted*, or (at the local level) imagine a being physically identical to me but with inverted conscious experiences. One might imagine, for example, that where I have a red experience, my inverted twin has a blue experience, and vice versa. Of course he will call his blue experiences "red," but that is irrelevant. What matters is that the experience he has of the things we both call "red"—blood, fire engines, and so on—is of the same kind as the experience I have of the things we both call "blue," such as the sea and sky...([Chalmers, 1996]: 100)
>
> ...as a *logical* possibility, it seems entirely coherent that experiences could be inverted while physical structure is duplicated exactly. Nothing in the neurophysiology dictates that one sort of processing should be accompanied by red experiences rather than by yellow experiences.([Chalmers, 1996], p. 100)

The possibility of inverted spectra has been thought about for a long time, and used in different ways, in the philosophy of language and mind. When it is used for different purposes, the details of the inverted spectra are not always the same. The key question is, what has to stay constant, while the subjective characters shift? When I was a graduate student in the 1960's, the key things that had to stay the same were the use of language and other observable behavior. Some of my teachers drew the conclusion that since changes in experience wouldn't show up in behavior, there was something fishy about experience; others drew the conclusion that the various forms of logical behaviorism were wrong.[2]

[2]See [Shoemaker, 1997] for the history of the argument as well as an extremely subtle analysis of its use against functionalism. At the end of his postscript, Shoemaker arrives

The latter use of the argument is legitimate and convincing. If behavior, including language use, is all that we hold constant across the individuals with different color experiences, it is clear that inverted spectra cases are possible, and to some extent no doubt actually occur. That there are individual differences in the color experiences sighted people have is clear from various forms of color-blindness, and the fact that color-blindness is hard to discover shows how easy it is for differences in color experience to be hard to detect at the level of language and behavior. That there are other individual differences, and that there might be a case in which things were perfectly shifted in some way, seems to me quite possible [Nida-Rümelin, 1997]. I am inclined to agree with Block that we "simply do not know if spectrum inversion obtains or not [Block, 1990/97]. (Shoemaker provides some reasons for thinking it does not in [Shoemaker, 1997].)

It does not follow from the success of these versions of the inverted spectrum argument, that a version of the inverted spectrum argument will be useful to Chalmers, for it does not follow that what Chalmers claims to be possible is possible. For Chalmers' purposes, what has to be held constant is not only the physical facts involved with language and observable behavior, and not only the functions of the color sensations, but *all* the physical facts that are in any way relevant to color experiences, down to the finest details of chemical processes in the rods and cones—the place where the differences in color experiences that we know of have their origin—and beyond, including experiences in the visual cortex and anywhere else relevant to vision and the experience of it. The plausibility of the inverted spectrum case in the context of an argument against logical behaviorism, simply does

---

at the conclusion that one can maintain a version of functionalism, in the face of the inverted spectrum argument, only by giving up a bit of common sense, that it makes sense to ask if your color experiences and mine are qualitatively the same. He opts to stick with functionalism and abandon that bit of common sense, where I would make the opposite choice.

not carry over to a case against antecedent physicalism. Thus, as with the Zombie case, we can grant Chalmers's the first requirement of his alternative possible world: we have twins with color-experiences systematically inverted relative to our own, and these inversions to not lead to any differences in linguistic or other behavior. But there is no reason to grant him the second requirement, that some of these worlds are physically indiscernible from our own. If the antecedent physicalist is right, there won't be.

We will return briefly to the Zombie and Inverted Spectrum arguments in Chapter 8, when we consider the purest form of the modal argument. There we will consider the framework of primary and secondary possibilities that Chalmers uses to present his argument. By that point we will be in a position to see how a certain resistance to the considerations presented in this chapter is built into that machinery and his use of it.

Chapter 5

The Knowledge Argument

Chapter 6

Recognition and Identification

Chapter 7

What Mary Learned

What then is Mary's *new* knowledge? The answer has to come at the level of the reflexive truth-conditions of her beliefs. In this chapter I'll apply the account developed in the last chapter to Mary's case, and argue that there is no problem there for the antecedent physicalist. Then I'll compare my view to the view of Laurence Nemirow and David Lewis, that Mary's new knowledge is a case of knowing how *as opposed to* knowing that. Finally I will look at a discussion between Paul Churchland and Jackson, which will enable us to see how the subject matter assumption underlies the knowledge argument. Just as epiphenomenalism is the real issue with the Zombie argument, I claim the subject matter assumption. Those who hold it, dualist or physicalist, have a problem with Mary's knowledge. Those who reject it, dualist or physicalist, do not.

7.1 Mary's new knowledge

Recall that the antecedent physicalist holds that there is a way of attending to a subjective character, that is only possible when one is having an experience of which it is the subjective character. There is a way of at-

tending to delightful aspects of the experience of eating of chocolate chip cookies, that is only possible when one is having that delightful experience. According to the antecedent physicalist, this does not mean that the state of enjoying a chocolate chip cookie is not a physical state, or that it could not be observed by others. Of course, there is no reason to suppose that observing my experience, perhaps by being a shrunken person inside my brain, would be itself particuarly enjoyable.

When we are attending to a subjective character in the subjective way, and wish to communicate what we are feeling or noticing, we use our flexible demonstrative, "this", as in "This feeling is the one I've been having". Let's label this use of "this" as a inner demonstrative: "this$_i$". Mary could use the following statement to express what she knew before leaving the Jackson Room, on the basis of her reading:

(1) Q_R is what it's like to see red.

and these statements to express what she learned upon seeing the ripe tomato:

(2) This$_i$ is what it is like to see red.

(3) Q_R is this$_i$ subjective character.

Let's call the beliefs expressed by (1), (2) and (3), b_1, b_2 and b_3. According to the antecedent physicalist the following can all be true:

- Q_R is a physical state, a physical aspect of the normal experience of seeing red;
- (1), (2) and (3) are true;
- When Mary leaves the Jackson Room she learns something new, by forming the new true beliefs b_2 and b_3, that she expresses with (2) and (3).

This new knowledge is a case of recognitional or identificational knowledge, as in the case with my new knowledge at the party with Dretske. We cannot get at it with the referential contents. We can get at the difference at the level of reflexive content. Let's look closely at b_1, b_2, and b_3.

The reflexive truth-conditions of b_1, the belief she had in the Jackson Room, and expressed with (1), are something like:

> b_1 is true iff the origin of Mary's Q_R concept, the concept involved in b_1, is the subjective character of the experience of seeing red.

b_1 was a detached belief when Mary got it, from reading a book; it never was connected to an act of attending to a subjective character. It is analogous to my first belief about Dretske, which existed for years before I had the opportunity to perceive Dretske himself, and which was connected to Dretske though a chain of communicative links. So to Mary's concept is the end of a chain of communicatives links; she formed the concept reading about Q_R in a book; the chain goes back to those who introduced the term, some of whom will have done on the basis of being subjectively aware of the sensation of red.[1]

The belief b_2 is analogous to my belief after Dretske introduced himself. That belief was attached to a perception of mine, which was of Dretske. Mary's b_2 is attached to an act of attention, which is an attending to of a certain subjective character. The referential truth conditions of (2) are exactly the same as (1). The reflexive truth-conditions of b_2 are different:

> b_2 is true iff the act of inner attention to which it is attached, is of the subjective character of the experience of seeing red.

[1] Note to Self: Here I am using Q_R as a name; need to check whether this is consistent with what was said about it before and what is said about it later.

Finally we come to b_3. This is the belief Mary expresses with (3), and it is the belief that Jackson found problematic. It is a belief about what people in general experience when they see red things, and it seems like the sort of thing she should have known in the Jackson Room, if she really knew all of the physical facts about color and color perception. The referential content of (3) is the same as that of (1) and (2). But the reflexive content differs:

> b_3 is true iff the act of inner attention to which it is attached, is of the origin of Mary's Q_R concept.

This is the new truth condition on Mary's beliefs, that results from the change that occurred when she saw the Tomato, and learned what it was like to see red. As in the cases of Larry, Garry and Terry, the change in Mary's beliefs does not result in any new on the truth of her beliefs *given* what they refer to. But it does impose new conditions on the truth of her beliefs, abstracting from what they refer to, the condition that the subjective character which is the origin of her old concept is also the one to which she is attending.

That's my account of how what Mary learned. But let's pause for a moment to make another point, while Mary's situation is before us. Take Mary back to the Nida-Rümelin Room for a moment. While there she had two concepts of the sensation of red, $Qwow$ and QR. They were unlinked. Not that it would have been pefectly coherent for her to have supposed, at that time, that

> (4) Q_{wow} is not Q_R.

The referential content of (4) is a contradiction. Since Q_{wow} and Q_R are one and the same subjective state, and there is no possible world in which that

one thing is not identical with itself, (4) cannot be true. But the reflexive content of (4) is a contingent proposition, roughtly that the subjective character that Mary experienced when she looked at a certain part of the plaid wall-paper, and of which she has certain memories, is not the same as the one that is the origin of the concept she acquired in the Jackson Room. We can say that (4) is conceivable for Mary, when she is in the Nida-Rümelin Room, because its reflexive content is consistent with the reflexive truth conditions of her beliefs. When Mary takes her next step, and sees the ripe tomato, and her beliefs change as above, she will realize, since she will recognize that wow is red, that

(5) Q_{wow} is Q_R.

At that point, (4) will no longer be conceivable for her. That is, the space of what is conceivable for Mary will have changed, as a result of her new knowledge. What is conceivable for Mary, now coincides with what is possible. Notice that there is no way Mary could have taken this step *apriori*.

To review. The antecedent physicalist holds that the subjective characters of experience are physical aspects of experiences, which we are able to attend to when we have those experiences, in a way that we cannot do so when we do not have them. Given that subjective aspects are physical aspects, they can be in principle observed, discussed, and written up in text books. So people can learn about subjective characters, on this view, which they have never had. They can know, of subjective characters they have never experienced, that they are the subjective characters normally associated with certain kinds of experience, such as seeing red. All of this does not mean that the antecedent physicalist needs to deny that when a person in this position, such as Mary, learns something new when they do finally experience the subjective character in question. The new knowledge,

as in the case with recognitional and identificational knowledge generally, is found at the level of reflexive content.

7.2 Recognitional Knowledge and Know-How

Laurence Nemirow has claimed, against the knowledge argument, that knowing what it is like is a species of knowing how [Nemirow, 1979, Nemirow, 1980, Nemirow, 1989]. Mary does acquire new knowledge, but it is not knowledge of a fact, hence not knowledge of a new fact, hence not an argument for non-physical facts. It is a matter of know-how. Mary learns how to recognize red things by sight, and how to recognize when she is having a red experience, how to imagine seeing red things. Nemirow's ability analysis has been adopted and defended by David Lewis [Lewis, 1990/97]. There is a very close connection between know-how and reflexive knowledge. In this section I'll explore how Mary's new knowledge relates to her new know-how.

To discuss know-how, we need to develop a couple of concepts from the philosophy of action.[2] I'll use "act" for particular events and "action" for types. So acts involve an agent performing an action at some particular time and place. Actions I'll divide into *accomplishments* and *executions*. Executions are identified and individuated by the particular movements involved. Accomplishments are identified and individuated by the results they bring about. So by moving my fingers (executions) I bring it about that the keys on my computer are depressed (accomplishments). By bringing it about that the keys are depressed, I bring it about that the state of the computer changes in certain ways; by doing that I bring it about that letters appear on the screen, and so forth (more and more accomplishments). Action is a matter of executing movements that have results; intentional action is a matter of executing movements for the purpose of

[2] See [I&P&T, 1993] and [Goldman, 1970].

getting results; successful action is a matter of executing movements that get the intended results.

A given action, execution or accomplishment, may constitute a *way of* bring about an accomplishment in certain circumstances. Depressing the keys is a way of making the letters appear on the screen *if* the computer is plugged in, the wires are intact, the right software is loaded, and so forth and so on.

In order for an action to be properly motivated by an agent's beliefs and a goal, the beliefs should close the gap between the action and the goal. That is, if the beliefs are true, the action should be a way of bringing about the goal. This will in general require two kind of beliefs: beliefs that in certain circumstances the action is a way of accomplishing the goal, and beliefs that those circumstances obtain. My moving my fingers is motivated by my goal of making letters appear on the screen. I believe that moving the fingers is a way of making letters appear on the screen, when the computer is plugged in, turned on and working properly, and I believe that it is plugged in, turned on and working properly. So my goal motivates my action.

I regard Know-how as a special kind of knowledge of "way-of" relations. (I'm also perfectly willing to talk about belief-how, which is a state that is internally like know how, except the way-of relation doesn't hold.) A more natural way to say what I said in the last paragraph is that I know how to make letters appear on the screen if the computer is plugged in and etc. But not any true belief about a way-of realtion constitutes know-how.

To know how to ride a bike, is to know which movements are a way of moving the bike in the direction you want to go without falling.[3] But not just any kind of knowledge of this will do. I may tell my wife that if

[3]For safety's sake, one might want to include something braking techniques too, but I'll ignore that here.

she simply turns gently in which ever direction she is starting to fall, while continuing to look in the direction she wants to go, she will remain upright and can go wherever she wants. She may believe me. That doesn't mean she knows how to ride a bike. Know how is a matter of attunement to a method, not possession of a formula describing the method. My granddaughter senses when her bike is falling in a certain direction, and turns gently into the fall. She has no idea that she is doing it; she couldn't say what she does.

Let's identify a method for bringing about R with the fact that an execution of movements M is a way of bringing about R in circumstance C. Know-how is a positive doxastic attitude —i.e. something belief-like, if not paradigmatically belief — towards a method, in which the movements are represented in a way that the agent can execute at will in a broad range of circumstances. This means the agent may not be able to name or describe the actions, but can probably demonstrate them. One reason that I want to regard this as a species of belief and knowledge, is that it seems to me that the fact that a certain type of execution will in certain circumstaces be a way of bringing about a certain result is something that is internally represented, and naturally regarded as a part of various concepts we have of various actions. Part of my concept of walking is that it done in a certain way, which I can demonstrate much more easily than I can describe. A second reason is that it seems to me that it is best to regard all of our knowledge as potential know-how; that is, our detached knowledge is of value only because in certain cirucmstances we can re-identify the objects it is about, and will then know how to do things vis a vis those objects, that we wouldn't know how to do otherwise. Recall the example of Krista Lawlor. I left the party knowing something about her, her name and some of her interests. The value of that was that combined with more

basic know how it enabled me, next time a saw her, to greet her by name, and ask something intelligent.

Thus there is a very close connection between recognition and know how. Recognition extends know how. When I realize that person A who plays role R in my life is also person B, then I learn that doing a certain thing to or for person A is a way of doing it to or for person B. I know how to talk to the person on the other end of the phone (talk into the end with the cord coming out of it). When I learn that you are the person at the other end, I know how to talk to you.

Suppose now that my sister teaches me to make a certain Aikido move, the Lotus Gives Birth, perhaps.[4] I finally get the idea. I cannot describe it in words in any very coherent way. And I quickly forget the name. But I do remember how to do it. I can demonstrate it (in the living room, slowly) and actually do it (on the mat, with lightning speed). It seems to me that the distinction between know how and knowledge that becomes a bit thin here. Suppose that my friend David, having read and memorized the Aikido book can give an excellent verbal description of the movements required for executing the Lotus Gives Birth, but cannot do it. He would have knowledge that a certain series of movements is the way to do Lotus Gives Birth, and knows how to describe them, but not how to do them. I know how to do them, but not describe them. But both of us have in our minds some representation of the movements; both bring the Aikido technique in question under a concept. Mine is an *executable* representation, like a schema, while his is not. From the point of view of the picture of cognition developed here, one might naturally say that knowledge-that is a certain kind of knowledge-how, knowledge that involves concepts that one can express verbally.

[4] See http://www.aiki.com.

Note that David and I could disagree about the right way to execute of Lotus Gives Birth. He could, and no doubt would, object to my demonstration as faulty, based on his more descriptive and theoretical knowledge. I might tell him his description must be wrong. One of will turn out to be right, the other wrong.

We do things with our minds, not just our bodies. There are mental actions we can execute at will — not very happily called "movements". One of them is attending to an experience we are having; another is trying to focus on what an experience is like so as to remember it; another is to focus on what it is like so as to recognize it. These are things we know how do do, with respect to experiences that play a certain role in our life; that is: the ones we have. I can't focus on what the experience of seeing red is like if I'm not seeing red, any more than I can shake hands with Fred Dretske when he is in North Carolina and I am in California.

Now let's consider Mary. When she is in the Jackson Room, she knows a lot about Q_R. But she doesn't know how to imagine being in Q_R, she doesn't know how to recognize Q_R in the way most of us do, and can't recognize red things in the way most us do. When she finally sees the ripe tomato, she will gain that know-how.

This may require a bit of effort on her part, however. I have seen puce many times, and been told that it is puce, but I cannot now recognize puce things on sight, and I couldn't tell you if I was having a puce sensation or not. I need to focus on the experience of seeing puce next time that I have it. Perhaps it takes an unusually lazy person to notice the effort involved in such a simple thing; philosophy needs all types. But I suspect only with a little effort will David Lewis know how to discriminate vegemite from marmite, should he ever be willing to ruin his example by trying them. Almost anyone who attends a wine-tasting seminar, in order to learn how

to no longer be satisfied with wine he can afford, will find it takes effort and practice to discriminate among one's sensations in the way the experts do.

One key to learning to recognize sensations is to engage our memories and imaginations at the time we have the experience. There is a wide range of cases. Color sensations are probably among the easiest for normally sighted people to imagine, recognize and remember names of. In the case of smells, we are likely to be much better at remembering whether we liked it or not, than being able to reproduce it in the imagination the way we can with colors. In all of these cases, there seems to be a phenomenon of attending to the experience, noticing things about it, including one's own reaction, the situation in which it arises, and so forth. That is, it seems that one is bringing the experience under concepts, including concepts like, "smells like *this*" where the "this" does not refer to the sensation or experience itself, but our reproduction of it in memory and imagination—not the impression, but the idea.

The conception of knowledge I have developed exalts knowing how, in that it insists that complete knowledge is tied to buffers that are ties to epistemic and pragmatic methods. In that context, it is easy to agree with Nemirow and Lewis that Mary's new knowledge is a case of knowing how, but not easy to agree that it is not a case of knowing that, or that all cases of learning from experience are so closely and effortlessly related to knowing how as is a case like Mary's.

Loar and Lycan give various reasons for preferring an account of the sort developed here. One is that Mary could have thoughts like, "If apples hadn't looked like $this_i$, I would have found them more attractive. She can retain this thought in memory, thinking of the look in terms of the experientially based concept of what red things look like. Another is that we can

apply our experience based concepts of subjective characters to other people. Mary can wonder if Harry prefers the look of apples to that of oranges, which she finds more attractive, because he actually has the experience she has of red when he sees orange [Loar, 1990/97], [Lycan, 1990].

7.3 Lewis and Eliminating Possibilities

In his essay "What experience teaches," David Lewis *defines* "phenomenal information" as irreducibly nonphysical ([Lewis, 1990/97]:583). Given this, he sees no hope for physicalism except to deny that there is phenomenal information. He sees the ability hypothesis he sees as the only alternative, and takes it to imply that phenomenal information is an illusion (593). I think this approach is unfortunate. The antecedent physicalist simply defines phenomenal information as whatever it is, if anything, that Mary learns, etc. That leaves us free to explore the phenomenon of phenomenal information and see if it involves anything nonphysical, and what its relations to gaining abilities might be.

The proposal I am putting forward may be an instance of what Lewis calls "The Fifth Way of Missing the Point". Lewis characterizes information in terms of eliminated possibilities. He says that there are conceptions of information that do not so characterize information. These conceptions foster "look alike" hypotheses:

> ...hypotheses which say that experience produces "information" which could not be gained otherwise, but do not characterize this "information" in terms of eliminated possibilities. These look-alikes do not work as premises for the Knowledge Argument. They do not say that phenomenal information eliminates possibilities that differ, but do not differ physically, from une-

> liminated possibilities. The look-alike hypotheses of phenomenal "information" are consistent with Materialism, and may very well be true. But they don't make the Knowledge Argument go away. Whatever harmless look-alikes may or may not be true, and whatever conception may or may not deserve the name "information," the only way to save Materialism is fix our attention squarely on the genuine Hypothesis of Phenomenal Information, and deny it.

The Fifth Way of Missing the Point involves appeal to the fact that Mary's mind has an internal structure of ideas for dealing with the world. Lewis assumes that appealing to changes in that structure to explain what goes on in Mary's case is simply to miss the point, by appealing to an irrelevant concept of information. He uses the analogy of taking a course in Russian, versus taking a course in English.

> Each of the look-alikes turns out to imply not only that experience can give us "information" that no amount of lessons can give, but also that lessons in Russian can give us "information" that no amount of lessons in English can give (and vice versa)..."

The subject matter assumption is apparent in Lewis's discussion. Loar says,

> Physicalists are forced into the Nemirow-Lewis reply if they individuate pieces of knowledge or cognitive information in terms of possible-world-truth-conditions...

To see Loar's point, recall the discussion of Mary in the Nida-Rümelin Room in the last section. For her, in the Room, it was conceivable that Q_{wow} and Q_R were different subjective characters. It became inconceivable when she

moved into the next room, and saw the ripe tomato. What she learned, cut down on what was conceivable. Of course, it did not cut down on what was possible, if we confine ourselves to the subject matter possibilities. But why should our conception of information be so inflexible as this? We'll return to these issues in the next chapter.

Let's consider Lewis's analogy for a moment. Suppose I am a native Russian speaker, taking a lesson on Cooking Pasta that is given in English. If my English is perfect at the beginning, and I have no knowledge of how to cook pasta, then we can characterize everything I learn in terms of the subject matter of the class. What is more likely is that I know something about Pasta, and have a partial grasp of English. Suppose, for example, that I know that "Vermicelli" and "Linguini" are both names of varieties of Spaghetti, and I have narrowed down the candidates for each to the same two varieties. In the course of the lessons, I will learn which variety each of the words stands for. This will eliminate possibilities, but they may not be subject matter possibilities. The teacher may have simply held up some Vermicelli and said, "This you boil only ten minutes, not twelve". I already knew that. What I learned was what that kind of spaghetti was called.

The line between learning about the meanings of English words and learning about Pasta is not a line between two concepts of information, one having to do with the elimination of possibilities, the other having merely to do with the presence of syntactic structures of some sort. The line is between what we took to be the subject matter of the class, and what we didn't. An English class for Russian Pasta Cooks might involve the very same words from the very same teacher; there the official assumption would be that the audience knows what kind of Pasta it is that the teacher holds up, and if they are too far away to see it they will know what kind it is when he tells them it should boil for only ten minutes. What they will learn,

when the teacher holds up the handful of Vermicelli and says, "Vermicelli cooks for only ten minutes" is the linguistic fact that "Vermicelli" stands for that kind of Pasta. But of course some of the audience members may be near sighted, and unable to remember how long different kinds of Pasta should boil, but may just happen to know what kind of Pasta "Vermicelli" stands for in English.

All of the contents that the content analyzer can find are contents that involve the elimination of possibilities. But the possibilities eliminated cannot all be represented permutations of the subject matter. Jon Barwise likes to say that language is a balancing act. What we may learn about from a particular utterance may be the context (if this is true, who must have said it? when must it have been said); the language (if this is true, what does "Vermicelli" stand for?) or the subject matter (if this is true, how long do you boil Vermicelli?). If we forget this, in looking at the knowledge argument, we will be caught between Jackson and Lewis, between misconstruing phenomenal information, and ignoring it.

The Russian example actually brings out the close connection between reflexive knowledge and abilities. Courses given in English and courses given in Russian presuppose quite different abilities on the part of the students. For most of us, knowing the meaning of the words of a language is not a matter of explicit beliefs about the words and their meaning. Rather, we have the ability to hear sentences in the language, combine the reflexive contents with other information, and form explicit beliefs about the subject matter.

7.4 Churchland's Challenge

I want to end this chapter by looking at an exchange between Paul Churchland and Jackson that helps show that it is the subject matter principle that leads to the problem. In the exchange, we see that Jackson thinks that physicalism is to be committed to this for some reason, while dualism is not.

Churchland tries a parity of reasoning argument, to show that there must be something wrong with the knowledge argument. As Jackson summarizes the argument, "Suppose Mary received a special series of lectures over her black and white television from a full-blown dualist," that gave her all the facts about dualism and qualia. "This would not affect the plausibility of the claim that on her release she learns something. So if the argument works against physicalism, it works against dualism too." ([Jackson, 1986]:569 summarizing [Churchland,1985].)

Imagine that dualism is true. There is no reason that Mary can't read about this in her room. And there is no reason that the subjective character of seeing red things can't be named Q_R and information about it printed in black and white, and given to Mary in her room. That is, imagine things are just as before, except that instead of being neutral between physicalism and dualism, the discussion of Q_R in Mary's texts emphasizes that it is not a physical state of the brain, but some kind of non-physical state.

Now imagine that Mary, having read and believed all of this, comes out of her room and sees a fireplug or a ripe tomato. It seems that there would be a experience gap. Mary could still think, "Ah, so *this* is what it is like to be in a brain state with that non-physical aspect I read about, the one that is involved in seeing red things, Q_R. There would still be a gap between Mary's reading about Q_R, and coming to know that it is the subjective

character of the experience of seeing red, and having the experience. And it seems as long as there is that gap, she learns something new when she has the experience. If it is a problem for the physicalist, shouldn't accepting dualism eliminate the problem?

Jackson replies that there is no reason to believe that everything about subjective characters could be told to Mary in the black and white room.

> To obtain a good argument against dualism...the premise in the knowledge argument that Mary has the full story according to physicalism before her release, has to be replaced by a premise that she has the full story according to dualism. The former is plausible, the latter is not. Hence, there is no "parity of reasons" trouble for dualists who use the knowledge argument ([Jackson, 1986]: 569).

Let the brain state that we go into when people with normal vision see red objects in normal light be called R. Take Q_R to be an *aspect* of R. Call the fact that S has aspect Q_R, *that-S-is-*Q_R. If Q_R is a physical aspect of brain states, then *that-S-is-*Q_R is a physical fact. If Q_R is a non-physical aspect of brain states, then *that-S-is-*Q_R is a dualist fact.

Now it seems that whether *that-S-is-*Q_R is a dualist fact or not, we can imagine Mary learning *that-S-is-*Q_R in the black and white room. She just reads a text book, written by an authoritative person, that says something like,

> There is a certain aspect of some brain states, that one is immediately aware of when one is in them, that we call their subjective characters. They are extremely important an interesting. One of the most studied subjective characters is Q_R, which is the subjective character of the experiences that normal people

> have when they see bright red objects, such as fireplugs or ripe tomatoes. For a long time it was not clear whether Q_R was a physical aspect of brain states or a non-physical aspect, but now it is known that . . ..

What is said up to the . . . would be agreeable, it would seem, to either an antecedent physicalist or a dualist. Once Mary reads that, she knows that Q_R is the subjective character of seeing red; that is, she knows *that-S-is-*Q_R.

Now, in either case, even though she knows *that-S-is-*Q_R, it still seems that intuitively she will learn something when she comes out of the Jackson Room and has the requisite experience of a fireplug or a ripe tomato. There will be a experience gap. No matter how carefully she has read the above paragraph—even if she has read whole books on Q_R, even if she has written them—it seems she will still be able to say,

> Oh, so this is what it is like to see red, that is, this subjective character is Q_R.

So it seems like the experience gap has nothing to do with physicalism. It seems like it is equally a problem for dualism and physicalism. The problem has to do with something that we said early in our description of Mary's situation, *that if something was known, it could be written down in black and white and she could read it in the room.* When Churchland assumes the very same thing for dualism, the experience gap problem emerges for it. And it is this step, that Jackson says isn't fair. Physicalist knowledge can be written down, dualist knowledge cannot be.

Now it seems to me that the true engine of the knowledge argument is coming to the fore. The physicalist is supposed to be committed to something about objectivity, that precludes Mary from learning about the

same physical fact in a new way, a subjective way, when she steps outside of the Jackson Room.

But my antecedent physicalist is committed to subjective ways of knowing physical facts in the following pretty clear sense. There is a way of knowing what an experience is like, that is available to a person who is having the experience, that is not available to others. A sighted person can know what it is like to see objects, in a way that a person who has never seen cannot.

Is there anything about this that violates the spirit of physicalism? I do not see that there is. All that is violated, is a false picture of knowledge. This is the view that there is some kind of knowledge that involves grasping a fact, not from any point of view—a view from nowhere.

This is a natural extension of the subject matter assumption. If the content of our beliefs is exhausted by the requirement their truth puts on their subject matter, then the methods of representation won't matter. The language won't matter, the context won't matter, and so forth. What is known will not constrain the knower to have any particular means of representation. Hence it will be possible to have any bit of knowledge, by means of representations that don't "locate" the knower in any way. This means that the references will not be by means of any roles that the subject matter plays in the life of the knower. The subject matter won't be, relative to the knower, I or you, this or that, here or there.

But this is a false picture of knowledge. A system of objective representation is a system for completing knowledge, and does not constitute the whole of knowledge. It would be, for us, like the phone book for poor Terry, a two chapters back, who cannot get a date. True knowledge is knowledge only because its potential for being attached to perceptions and actions.

Science is supposed to be objective in several senses. The experiments

should be replicable by different people in different laboratories. The observations should be public and checkable. The results should depend on what happens in the experiments and observations, not what a particular person, group, or funding agency wants to be true. And so forth. In science, as in all human communication, we seek an appropriate mode of representation. Scientific results should be published a journal and in a language that many scientists have access to; new terms should be explained in this well-known language; and information should be conveyed in ways that do not require the reader to know details of the writer's situation or personal circumstances that are not supplied. All of this does not add up to any special commitment on the part of scientists in general or physicalists in particular, to the subject matter assumption and the particular doctrine of objectivity it entails. But without the subject matter assumption the knowledge argument is no more of a problem for the physicalists than it is for the dualist; with the assumption, it is a problem for both.

Chapter 8

The Modal Argument

> ...it is downright self-contradictory to say (in a reasonably constructed and interpreted language) that Smith is Jones, or that I am you. The Mont Blanc cannot conceivably be identical with Mt. Everest!
>
> [Feigl, 1958/67]: 62

The Zombie argument we examined in Chapter 2 is a modal argument. It is claimed that a something is possible, a world physically indiscernible from ours, but with no consciousness. From the existence of a possibility, an inference is made about the actual world: physicalism is false. I maintained that the argument did not work against an *identity* theory of subjective characters.

The first and simplest modal argument was advanced against such an identity theory, however, by Saul Kripke in *Naming and Necessity* [Kripke, 1980/97]. A version of this argument is also put forward by Chalmers. In this last chapter I'll examine these arguments, and then close by saying a bit about the "explanatory gap".

8.1 Kripke's Argument

Consider Q_R. Suppose its physical correlate has been identified. Many scientists, including Mary, as she published articles while trapped in the Jackson Room, thought that the correlate was B_{47}, that is, the brain state with the scientific-structural description that we will imagine to be conveyed by "B_{47}". But it fact it turned out to be B_{52}. The antecedent physicalist now claims they are just one thing, one and the same property or condition. If they are one thing, then there is no way they can be two things, and there is no possible world in which "they" occur separately. In maintaining identity, then, the antecedent physicalist maintains necessary identity. So, conversely, if there is such a possibility, Q_R and B_{52} are not one thing, and the antecedent physicalist is wrong. The modal argument claims that there seems to be such a possibility, and that it cannot be explained away.

Kripke's argument is recognizably a descendant of the experience gap argument, but more powerful because it draws not only on our intuitions, but also on the a framework for discussing issues of necessity and possibility developed by Kripke and others over the past forty years. The argument focused on pain, and assumed that the identity theorist claims it to be identical with stimulation of C-fibers ([Kripke, 1980/97]: 446ff.) If the identity is true, it is necessary. There is just one thing, one property, that is both C-fiber stimulation and pain. This means that there could not be a C-fiber stimulation that was not a pain, nor a pain that was not a C-fiber stimulation. This is surprising, Kripke says, but not yet fatal to the identity theorist, for, perhaps the identity theorist can show,

> ...that the apparent possibility of pain not having turned out to be C-fiber stimulation, or of there being an instance of one of the phenomena which is not an instance of the other, is an

> illusion of the same sort as the illusion that water might not have been hydrogen hydroxide, or that heat might not have been molecular motion...([Kripke, 1980/97]: 447).

In these cases, the key fact is that the designators "water" and "heat" designate contingently. This gives a statement like "heat is molecular motion" an "illusion of contingency". It creates the possibility of someone being in "qualitatively" the same epistemic situation.

In my terminology, the point is that the referential content of statement may be a necessary proposition, but there may be other salient contents, that are contingent, and that provide the sense of contingency. If we take the reference of the terms in "Heat is molecular motion" as given, then our content is necessary. But given only that "Heat" refers to the process that causes certain sensations in us, we have the attributive content that molecular motion causes those sensations, which is contingent. If we think it is contingent that heat is molecular motion, that is probably the contingency we in some sense have in mind.

Now suppose that "pain" worked like we assumed that "heat" did. Just as heat is the cause of heat sensations, pain is the cause of pain-sensations. Then the identity theorist could say something quite analogous to what was said about the heat case in the last paragraph. Taking the reference of "pain" and "vibrating C-fibers" as given, we have a necessary proposition. But given only that "pain" refers to the *cause* of pain-sensations, we would have a contingent proposition, to the effect that vibrating C-fibers causes those sensations.

But Kripke says, quite correctly, that this avenue is not open to the identity theorist. There simply is not a distinction between pain and pain sensations, as there is between heat and heat sensations. This is a point we emphasized in Chapter 3. He makes the point in various ways, however, to

some of which I will take exception. The following remark seems exactly right:

> In the case of molecular motion and heat there is something, namely, the sensation of heat, which is an intermediary between the external phenomenon and the observer. In the mental-physical case no such intermediary is possible, since here the physical phenomenon is supposed to be identical with the internal phenomenon...

Here Kripke is clearly thinking about *having* pain; he is, in Hume's vocabulary, talking about the impression. But elsewhere he talks about "picking out" pain. This requires something that is doing the picking out. "Picking out" suggests either an idea or a term that *refers* to pain, or someone who *recognizes* pain. In either case it is not only pain that is involved, but our concept of pain; not just the impression, but also the idea. What he says does not seem correct:

> Pain...is not picked out by one of its accidental properties, rather it is picked out by the property of being pain itself, by its immediate phenomenological quality.
>
> If any phenomenon is picked out in exactly the same way that we pick out pain, then that phenomenon *is* pain. ([Kripke, 1980/97]: 448)

Impressions and Ideas

While it is quite right that pain is not something we feel of because of an intermediate, contingently connected, appearance, it does not follow that no contingently connected intermediaries are involved in thinking about pain.

We need to distinguish between our pain, and our idea or concept of it—between the impression and the idea, as Hume would put it. The first is a state that we are sometimes in. The second is a component of our thoughts and memories and anticipations of that state. Our concept of pain is not an intermediary between us and the experience of pain. But it is an intermediary between us and pain, in that it provides our way of thinking of pain when it is absent, and classifying it when it is present. How does that concept pick out pain?

Think of the successions of concepts that Mary has had of the subjective character of red. First, in the Jackson Room, it was known to her from a textbook. She certainly had, at that time, a concept that would pick out that subjective character in the sense of referring to it. This concept also incorporated knowledge of some paradigmatic red objects, such as ripe tomatoes, and so did enable her to pick out red and the sensation of red, in the sense of recognize, when she finally made it to the ripe tomato.

She had a second concept of red in the Nida-Rümelin Room, when she noticed several patches of red, and gave the color the name "wow", and the subjective character the designator Q_{wow}. This concept also picks out Q_R in the sense of being *of* Q_R. But it provides her with quite different tools for recognizing subsequent cases of wow and Q_{wow} than the concept she acquired in the Jackson Room. She will have memories of what it was like to see wow and be in Q_{wow}. She could close her eyes for a few moments and re-identify the patches, or pick out new ones if the patterns changed. But she has no knowledge incorporated into this concept about paradigmatically wow objects.

When she leaves the Nida-Rümelin Room, she bring both of her methods for picking out colors and sensations into play. She recognizes the sensation she has when she sees the ripe tomato as Q_R, because of her knowledge that

ripe tomatoes are red, and she recognizes it as Q_{wow} because of what it is like to be in it. From all of this she learns that Q_{wow} is Q_R and wow is red. She acquired new know-how; she can identify red when she sees it and Q_R when she is in it. She also knows how identify references to Q_{wow} in her text book, she learns that using the term "Q_R" is a way of referring to Q_{wow}. At the same time, as we noted, she has restricted the space of what she can coherently conceive; she now cannot imagine changing the color of a wow thing to red.

Now she is looking at the red tomato, and attending to the subjective character Q_R. Her concept of Q_R is attached to an instance of it. Her concept is not an intermediary between her and Q_R. She is in state Q_R, and the concept is not involved in that. She is attending to Q_R, and the concept is not an intermediary there, either. One has the ability to attend to the subjective characters of the states one is in; concepts are not required, although they may intrude.

Mary then has a number of ways of thinking of Q_R; as "this$_i$ subjective character" (in virtue of consciously attending to it); as "Q_R" (in virtue of its being the subjective character of red objects); as "Q_{wow}" (in virtue of its fitting her memories of that character). Each of these ways of thinking picks out Q_R, in two rather different senses. First, each of these concepts is *of* Q_R. Second, each is tied to a way of identifying or recognizing Q_R. The last two are tied to different methods of picking out, in terms of recognitional criteria. In the case of bare attention to her subjective character, there are no recognitional criteria, since she is not identifying the subjective character as anything beyond that to which she attends.

Now all of this is quite contingent. Q_R is one thing, the act of attending to it is another; Q_R is one thing, the role of being the subjective character of the experience of seeing red objects is another; Q_R is one thing, memories

of it are another. So there is no lack of contingency to serve as the material for illusions, or for non-referential contents, for Mary's thoughts about Q_R. None of these things, however, is an *appearance* of Q_R, none of them is a further sensation that is involved in the awareness of Q_R. Mary is not aware of Q_R *via* awareness of something else. It is a state she is in, and she can be aware of and can attend to it. She can also remember things about it and the situation in which it occurs, theorize about it, publish papers and books about it, and the like. While simply being aware of her sensation does not involve intermediaries, all of these activities do require various intermediaries: memories, concepts and words.

The Feigl Room

Now let's take Mary into the Feigl Room, where the awful truth is disclosed to her: Q_R simply is the brain state with scientific structural description B_{52}. That is, she is told,

$$Q_R = B_{52}.$$

Mary can hardly believe it. She had put her money, staked her professional reputation, on B_{47}, the older candidate for being Q_R that once seemed very promising. She looks over the data. She finally agrees. Surely she will have a vivid sense of contingency. She might have been right. That is, B_{47} might have been $this_i$; it didn't *have to be* B_{52}.

Given the picture of the content of thought developed in the previous chapters, we won't think of Mary's thoughts as merely having referential contents, and her thoughts as simply contradictory, with perhaps accompanying illusions. There is more flexibility to thought than that. She has a thought, "$this_i$ is B_{47}", the referential content of which is false, and necessarily so. So it's incorrect to suppose it might have been so, as she does.

She has another, "*this*$_i$ is B_{52}" the referential content of which is true, and necessarily so. So she is incorrect to suppose this might not have been so.

But each of these thoughts have many non-referential contents, which have a different modal status than the fully referential, subject matter content. For example, the first referentially contradictory thought has non-referential contents to the effect that the subjective character to which she is attends is B_{47}. This is a contingently false proposition, that might have been true.

To this one might reply, that that is a contingent proposition, but it is not what she was thinking, she was thinking about the subjective character she was experiencing, that *it* might have been brain state B_{47} instead of brain state B_{52}. But I agree that the necessarily false proposition was what she was thinking. I am claiming only that the *contingently* false proposition was *also* a truth condition of her thought; it was the content given meaning but not the reference of "this$_i$. Such non-referential contents are salient aspects of the thought, that can be used to explain the sense or as Kripke says "illusion" of non-contingency.

Kripke describes the strategy that works in the case of heat and molecular motion, water and hydrogen hydroxide, but not pain as stimulated C-fibers, as follows:

> The strategy was to argue that although the statement itself is necessary, someone could, *qualitatively* speaking, be in the same epistemic situation as the original, and in such a situation a *qualitatively* analogous statement would be false...

But in the case of pain and C-fibers, this won't work:

> To be in the same epistemic situation that would obtain if one had a pain *is* to have a pain; to be in the same epistemic situ-

> ation that would obtain in the absence of a pain *is* not to have a pain.

The idea is that we can imagine changes in the world that leave a subject with the same sensations, or perhaps more broadly the same evidence, while changing the facts that her language and thought refer to. But when the subject matter of the thought and language is the sensations themselves, there doesn't seem to be any room for maneuver.

Consider Mary just as she waits outside the Feigl Room. She doesn't know whether Q_R is B_{47} or not; she thinks it is, she has argued that it is, but she realizes the evidence is not conclusive. She stares at a red wall and hopes, "this$_i$ subjective character is B_{47}." What she is attending to is not B_{47}, but B_{52}. And we can't say that B_{47} might have caused that sensation, because it isn't *causing* that is at issue, it's *being*. That sensation *is* B_{52}.

For Q_R to be B_{47}, it would have to have the location, composition, and other factors that are built into our (pretend) scientific name, "B_{47}". Clearly, for all Mary knows, Q_R does have those properties, for that in fact is the hypothesis she thinks is most likely. There are, in fact, a number of situations that are compatible with Mary's epistemic situation, although we cannot get at them, with merely the referential contents of her thoughts. But are these possible situations? After all, it is essential to B_{52} to have just the location, composition and other characteristics that are incorporated into its scientific name.

Mary is a trained scientist and has three concepts, being Q_R, being B_{47} and being B_{52}. In her trained scientific mind there is a plausible case for identity between the referents of the first two concepts, but no more. Scientists must always accept the possibility of being wrong. So from her point of view there are two live possibilities, or conceivabilities. One she would represent by linking the concepts Q_R and B_{47} and express by saying

"Q_R is B_{47}". The other she would represent by linking the concepts Q_R and B_{52}, and express by saying "Q_R is B_{52}". We saw in earlier chapters that we cannot capture the content of the change in belief except by retreating to the reflexive level. The same goes for understanding the nature of the possibilities that Mary contemplates. As she stares at the red wall, Mary has a hope, one of whose reflexive contents is that her concept B_{47} is of the subjective character of her current color experience.

Mary thinks about her current sensation in two ways, in contemplating the various scientific identities while she stares at the wall. One way, by attending to it, is not mediated by any appearance, description, or concept (though it may be attached to her concept of Q_R). The other is by way of individuating scientific properties, associated with the scientific terminology. This is not direct, but mediated by her concept, the language of scientists, the properties they associate with the language, and so forth. To find Mary's sense of contingency, to find something coherent that she may hope for even though, as we know, she is wrong, we retreat along the lines of the scientific description, not along the lines of her demonstrative thought about her own state.

The Autocerebroscope

Another surprise awaits Mary in the Feigl room: the autocerebroscope ([Feigl, 1958/67]:14, 14n). With this Mary can simultaneously have a sensation and observe it in her own brain, through the autocerebroscope. Feigl imagined this on analogy with a Fluoroscope, so that Mary would be looking at something like a pattern on a monitor of her brain activity. With our more up to date imaginations, perhaps we can imagine it attached to some sort of electron microscope, that can be aimed right at the location or locations in her brain relevant to her subjective character; or at least to the

places where activity would differ depending on whether B_{52} of B_{47} were occurring. B_{47}, she is told, is actually the subjective character associated with seeing puce objects. She learns quickly how to use the scope. She watches (with her right eye) what happens in her own brain as she shifts her look (with her left eye) from a red surface to a puce surface. There is no doubt about it. She was wrong.

Still, she could have the following thought (using "this$_{ac}$" for her attention to the autocerebroscope):

> *This*$_{ac}$ brain state (left eye on puce surface) might have been *this*$_i$ subjective character (left eye on red surface). I might have been right.

Again, there is no line of semantic retreat on the "*this*$_i$ subjective character" side of the identity. On the other side, the autocerebroscope provides Mary with as direct perception of a brain as we can imagine, almost as good as being inside Leibniz's mill-size brain, or in the boat with the tiny scientists of *Fantastic Voyage*. That is still not as direct, however, as *being in* a brain state. There is an appearance/reality distinction to be made. Mary could consistently imagine looking at the red surface with her left eye and having Q_R while looking in the autocerebroscope with her right eye and having the experiences she in fact has only when she looks at the *puce* surface with her left eye. If we abstract from the the reference of the autocerebroscope pattern to the brain state B_{47}, while retaining its association with the name "B_{47}" as it appears in Mary's thinking (and publications), we get roughly the proposition that Q_R appears like so-and-so on an autocerebroscope, is called "B_{47}", and is what I was referring to in my journal articles." That's the coherent content of Mary's hope and imagination, the coherent basis that provides an illusion of contingency for the awful neces-

sary truth that Q_R is really B_{52}, her conjecture wrong, her career shattered, and probably a long career in minor adminstrative posts the most she can hope for.

I think, then, there are enough contingencies, discoverable by using the content analyzer as we semantically retreat from one or the other terms in the various necessary identities we have considered, to explain Mary's various feelings of contingency in the face of them.

8.2 Primary and Secondary Possibilities

Chalmers' distinguishes between the *primary* and *secondary* intensions of a statement. In my terms, this is roughly the difference between the truth-conditions given the descriptive meaning, and the truth-conditions given the reference[1]. Suppose that "water" means "the watery stuff". Then the primary intension of "Lake Erie is full of water" will be that Lake Erie is full of the watery stuff; this will be true in worlds where Lake Erie is full of the main wet drinkable liquid in the world. The secondary intension will be that Lake Erie is full of H_20. This will be true in any world in which Lake Erie is full of H_20, even if it is not the predominant wet drinkable stuff in the world.

At the level of secondary intensions, the possible and the a priori do not coincide. There can be modal discoveries and surprises. It is necessary that water is H_20, but not something we could have known a priori. But, Chalmers says, there are no surprises or discoveries at the level of primary possibility (except discoveries based on conceptual analysis). The primary intension of "Water is H_20" is not necessary, and the primary intension of "Water is the watery stuff" is.

This apparatus gives Chalmers' a more direct way of advancing what he

[1] Note to self: get the terminology consistent here and in chapter on reflexive content

takes to be Kripke's main insight. This is that there is a property, which we can call "being painy", that provides us with a primary intension for statements with the word "pain" in them. The primary intension of "pain is painy" is a necessary truth. There is *no* physical state such that it is a necessary truth that *it* is painy. There are, Chalmers supposes, worlds in which stimulated C-fibers are not painy, for example. In fact, if stimulated C-fibers are the physical basis of pain in our world, his Zombie world is just such a world. For our Zombie-twins don't have any painy sensations, when their C-fibers are stimulated. Kripke's insight was really that Zombies were possible.

Now we said in Chapter 4, about his Zombie world, that it assumes what the antecedent physicalist denies, that the subjective character of pain is a physical aspect of brain states. If this is so, there will be no possible world in which the "physical basis" of pain obtains without being painy, because the paininess of the physical basis is one of its physical aspects. We can now elaborate on that response, using the materials of the chapters on the knowledge argument. We do, or might, have two concepts of the subjective character of pain, just as Mary did of the subjective character of red. One is drawn from our experience of pain. We have a concept of pain that is involved in our thinking about pain, remembering pain, anticipating pain, writing about pain, and doing research on the physical bases of pain. We have another concept of pain (let us suppose) as the stimulation of C-fibers. But these are, in fact, the same property. We can have two concepts of the same property, just as we can have two notions of the same person.

We saw that to get at the knowledge one gains, when one recognizes or identifies an individual, or when one recognizes or identifies a universal, one needs, or may need, to get at the epistemic change in terms of reflexive content. That is because the change is not a change in the properties of

the subject matter, but a change in the way the system of representation is structured.

Loar developed an account that uses this approach; Loar argues that there can be two predicates or concepts of the same property, one introduced through a process of inner demonstration, the other scientific [Loar, 1990/97]. In this case, the two concepts will not have different primary intensions. In response to this Chalmers says,

> But how can two primary intensions coincide without our being able to know it *apriori*? Only if the space of possible worlds is smaller that we would have thought *apriori*. We think the intensions differ because we conceive of a world where they have different reference, such as a zombie world. Loar's position therefore requires this world is not really possible, despite the fact that we cannot rule it out on conceptual grounds...

Chalmers simply isn't facing up to the human condition, or, as one might put it, the limits of pure imagination. We can create reflexive possibilities through thought and language, but not real ones. Julius walks by. I don't recognize him. I have a perceptual buffer, unattached to my Julius file. Is that fellow Julius? Maybe so, maybe not. It could have been a lot of people. Could have been Dan Flickinger, or Julius, or maybe Pierce Brosnan. Those are all real epistemic possibilities, or as I shall say, *conceivable situations* for me. They are possibilities concerning how my system of representations fits onto the world. *A priori reflection will not make them disappear.* Given the way the system does in fact fit, there is one necessity, that that man is Julius, and two impossibilities. On any reasonable epistemology of universals, of properties and relations and states and subjective characters, the power of *apriori* reflection will be similarly limited.

The answer to the question that is the first sentence of the quote from Chalmers, "...how can two primary intensions coincide without our being able to know it *apriori*?" is basically: unreflected identity. If we do not know that Hesperus is Phosphorus, then we will represent the world as having more possibilities than it does. If we don't know that a fortnight is two weeks, we'll represent the world as having more possibilities than it does. And if we don't know that having pain is having vibrating C-fibers, or that red is wow, or that Q_R is B_{52}, we will represent the world as having more possibilities than it does.

8.3 Reflexivity and Indexicality

Chalmers system of primary intensions, as I have described it so far, can't deal with indexical statements and the thoughts that they express. Return to the case of Dretske and I at the party. Consider the statement, "I am talking to you." The secondary intension of this will be the set of worlds in which John Perry is talking to Fred Dretske at the time of the party. That is not any kind of necessary truth. But it seems like there ought to be something like a primary intension, that is necessary truth or something close to it, roughly corresponding to what I call the truth conditions of the utterance given only the meaning but not the contextual facts.

Chalmers gets at this primary intension by using *centered* worlds rather than ordinary worlds for his primary intensions. A centered world is a world plus an agent and a time. The primary truth conditions of a statement is,

> ...a set of centered possible worlds in which the statement, evaluated according to the primary intensions of the terms therein, turns out to be true. The primary truth conditions tell us how the actual world has to be for an utterance of the statement

> to be true in that world; that is, they specify those *contexts* in which the statement would turn out to be true. (63)

The primary intension of "I am talking to you" will be the set of centered worlds in which the agent is talking to the person the agent is talking to at the relevant time. That will not include all centered worlds, but it will include all of them in which the agent is talking to someone. So we get something that is conditionally necessary, and quite different from the set of worlds in which John Perry is talking to Fred Dretske at the time of that party.

A primary intension is a property or condition that is associated with a term, and provides a condition that an object must satisfy, to be designated by the term. But not all terms provide such conditions; the term "you" does not. What it provides is a *role*, that an object must play relative to the speaker at a time. To get a primary intension, we need a speaker and a time, in addition to the role, and that is what the centers provide.

Now let's go back to me at the party. I had no doubts about who I was. But I didn't know who the person I was talking to was. How do we get at the possibility that I took to be true, that I was talking to someone other than Fred Dretske? That would be the set of centered worlds in which I am the agent, the time of the party is the time, and I am talking to someone other than Dretske. Conversely, suppose that I knew who Dretske was, but had forgotten who I was (serious epistemology sometimes has that effect on me). That possibility would be represented by the set of centered worlds in which the agent of the center, at the time of the party, is talking to Dretske.

In these cases, there seems to be a pattern about our thoughts about what is possible and what is not, fact that is captured by the centered worlds, that we cannot capture with uncentered worlds. By shifting to centered worlds, Chalmers allow us to get at this pattern, but at the same

time he undermines the claim made above in his response to Loar, that at the level of primary intension, our apriori imagining and reasoning is an infallible guide to possibility.

That pattern is this[2]:

1. When a mental or linguistic term is associated with an role R, rather than a property, there is a gap between the role provided and the primary intension needed. (E.g., "I", "now", "you".)

2. Where X is a primary intension that picks out the person or thing playing role R, the agent may think something referentially impossible: *R is not X*, and be unable to discover the mistake on conceptual grounds, by reasoning *apriori*. (E.g., I might have thought, "You are not Fred Dretske" while talking to Fred Dretske.)

3. There will be no subject matter possibility corresponding to this thought. This is, it is not possible for the objects that the agent is actually thinking about to be different, if they are identical.

4. The underlying possibility will involve something *else* playing the role. (The worlds in the primary intension will be centered worlds, in which I am the agent at the center, and I am talking to people who are not Fred Dretske.)

It is just this pattern that the antecedent physicalist sees in the case of Chalmers thought experiments:

1. Our experiential concepts of experience types are tied to roles R they play in our lives; *this$_i$ experience* (the one I am having); *this$_m$ experience* (the one I am remembering). Ie, these concepts provide only roles, not primary intensions.

[2]See my essay [Perry, 1977].

2. Where X is a primary intension picking out the state that actually plays this role, the agent may think something that is referentially impossible: *R is not X*. (E.g., "This sensation is not stimulated C-fibers", "Q_R is not B_{52}")

3. There is no subject matter possibility corresponding to the thought.

4. The underlying possibility is of some *other* state occupying that role. (Worlds in which I am attending to some other than stimulated C-fibers; worlds in which Mary is attending to B_{47}.)

Chalmers comes up against this possible reply of Loar's and others who share his approach in various footnotes. The response is always to deny 3; this is to insist that the Zombie thought experiment shows that the possibility is real, and that therefore this pattern does not apply. From the point of view of the antecedent physicalist, that is begging the question.

The Zombie thought experiment is *conceivable*, for dualists who do not believe that phenomenal states are brain states. There is no internal coherence, in a world in which their representations of phenomenal states are not connected to the same properties as their representations of brain states, and so no internal coherence in an alternative world in which no properties occur to which their phenomenal concepts refer, but properties to which their brain-state concepts refer do occur. That is, the thought "Brain states are not phenomenal states" is conceivable for them, because its reflexive content — the content given the internal structure of their thoughts, but abstracting from reference—is possible. For the antecedent physicalist, who believes in identity, the same thought is not conceivable, for the reflexive content of the thought "Brain states are not phenomenal states" is not possible. In the same way, the poor history student can easily conceive that Cicero was not Tully, but her Professor cannot. The dualist

can certainly check on conceivability *apriori*. But is the world, which the dualists conceive, *possible*, as opposed to merely conceivable for them? That is, is the referential content of the thought "the brain states occur without the phenomenal states" possible? This cannot be determined *apriori*, any more than one can show that Cicero might not be Tully by performing poorly on a Roman History exam.

8.4 Categorical denials of identity

Mary's thoughts, however, differ from those Leibniz or Ewing might have, presented with the alleged identity between Q_R and B_{52}, or given a chance to use the autocerebroscope. At least as I have embellished her biography, Mary had no problem with the idea of a true identity between brain state and subjective character, merely thoughts, hopes, and disappointments about which particular ones happened to be true. For Leibniz and Ewing —and Kripke too, I think—very very idea is absurd. Such philosopher would have the thought, looking through the autocerebroscope,

> $This_{ac}$ brain state is not, and could not be, $this_i$ subjective character.

According to the antecedent physicalist, these philosophers are not only wrong, but necessarily so; the brain state not only is the subjective character, it could not be other. We could, however, find suitably non-referential consistent backup contents to explain their sense of contingency, perhaps that there is a state, that the autocerebroscope image is of, that is correlated with, but not identical to, Q_R.

But what can the antecedent physicalist say about the motivation for their denial? One reason might be that it is very odd that it is like something to be in certain brain states, and would be even odder if this were a

physical aspect of the brain state, rather than some non-physical property. I am sympathetic with the idea that it is very odd that it is like something to be in certain brain states. The world is a very odd place, at least for the philosopher, and this is one of the leading oddities. But it is certainly *not* odd in the sense that there is some *other* kind of state, such that it wouldn't be odd, or would be less odd, if it were like something to be in those states. Simply to say "non-physical" is not to provide a less odd kind. If one puts some content into the idea of non-physical by giving examples, such as the state of pain, or the subjective character of seeing red, then one begs the question.

The only other reason I can think of, is simply the Ewing intuition, now deprived, I hope, of the power of modern analytical philosophy to support it . What it is like to have an experience, is nothing at all like what it is like to look through an autocerbroscope at one's own brain states, or to be miniaturized like the folks in *Fantastic Voyage* and look at the brain states of other. But why should it be? Why should the one experience (being in a certain brain state) be like the other (being in the brain states one is in when one (somehow) looks at that first experience)?

Chapter 9

Bibliography

Bibliography

[Barwise & Perry, 1983/99] Barwise, Jon and John Perry. *Situations and Attitudes* (Stanford: CSLI Publications, 1999); reprint, with additions, of Barwise, Jon and John Perry. *Situations and Attitudes* (Cambridge: Bradford-MIT, 1983).

[B&F&G, 1997] Block, Ned, Owen Flanagan and Güven Güzeldere, editors. *The Nature of Consciousness* (Cambridge, Mass.: Bradford-MIT, 1997).

[Block, 1990/97] Block, Ned. Inverted Earth. In [B&F&G, 1997]: 478-693. Reprinted from James Tomberlin, ed., *Philosophical Perspectives*, Vol. 4 (Atascadero: Ridgeview Publishing, 1990): 31-52.

[Block, 1995a] Block, Ned. On a Confusion About a Function of Consciousness. *Behavioral and Brain Sciences*, 18 (1995), 227-247; reprinted in [B&F&G, 1997], 375-415.

[Block, 1995b] Block, Ned. Mental Paint and Mental Latex. in [Villanueva, 1995].

[Block, forthcoming] Mental Paint. In [Hahn & Ramberg, forthcoming].

[Block, tbf] His best paper for anticipation of Chalmers quoted distinction. Search Chalmers notes.

[Chalmers, 1996] Chalmers, David. *The Conscious Mind*. (New York: Oxford University Press, 1996).

[Churchland,1985] Churchland, Paul. "Reduction, Qualia, and the Direct Introspection of Brain States," *Journal of Philosophy*, LXXXII (1985): 8-28.

[Churchland, 1989/97] Churchland, Paul. "Knowing Qualia: A Reply to Jackson." In [B&F&G, 1997]: 571-577. Originally Chapter 4 of [Churchland, 1989]: 67-76.

[Churchland, 1989] Churchland, Paul. *A Neurocomputational Perspective*. Cambridge: Bradford-MIT, 1989.

[Clarke-Collins, 1711ff] The Clark-Collins Controversy.

[Crimmins & Perry, 1993] Crimmins, Mark and John Perry. The Prince and the Phone Booth. *Journal of Philosophy*.XXX Reprinted in [Perry, 1993].

[D&H, 1993] Davies, M. and G. Humphrey. *Consciousness*. Oxford: Blackwells, 1993

[Dennett, 1988/97] Dennett, Daniel C. Quining Qualia. In [B&F&G, 1997]: 619-642. Reprinted from *Consciousness in Contemporary Science*, edited by A. Marcel and E.Bisiach. (Oxford: Oxford University Press, 1988): 43-77.

[Dodwell, 17??] Dodwell's book vs. Locke.

[Evans, 19??] Evans, Gareth. The Causal Theory of Names.

[Ewing, 19??] Idealism? Or Where?

[Farrell, 1950] Farrell, B. Experience. *Mind*, vol. 59, 1950: 170-98.

[Fleischer, 1966] Fleischer, Richard, Director. *Fantastic Voyage.* Starring Stephen Boyd, James Brolin, Arthur Kennedy, Edmond O'Brien, Arthur O'Connell, Donald Pleasence and Raquel Welch.

[Feigl, 1958/67] Feigl, Herbert. The "Mental" and The "Physical". In [Feigl, 1967] Reprinted from [Feigl, 1958b].

[Feigl, 1967] Feigl, Herbert. *The "Mental" and The "Physical": The Essay and a Postscript.* Minneapolis: University of Minnesota Press, 1967.

[Feigl, 1958b] Feigl, Herbert, Michael Scriven and Grover Maxwell, *Minnesota Studies in the Philosophy of Science*, volume II. Minneapolis: University of Minnesota Press, 1958.

[Frege, 1892] Frege, Gottlob. Über Sinn und Bedeutung. *Zeitschrift für Philosophie und Philosophische Kritik*, NF 100, 1892: 25-50. Reprinted in Frege, G., *Funktion, Begriff, Bedeutung: Funf logische Studien.* Edited by Günther Patzig. (Göttingen: Vandenhueck & Ruprecht, 1980): 40-65.

[Frege, 1892/1960] Frege, Gottlob. On Sense and Reference. Translation of [Frege, 1892]. In *Translations From the*

Philosophical Writings of Gottlob Frege. Edited and translated by Peter Geach and Max Black. Oxford: Basil Blackwell, 1960: 56-78.

[Goldman, 1970] Goldman, Alvin. *A Theory of Human Action* Englewood Cliffs N.J.: Prentice Hall, 1970.

[Hahn & Ramberg, forthcoming] ESSAYS ON TYLER BURGE, EXACT TITLE UNKNOWN. (Cambridge: MIT, forthcoming).

[Hampshire, 19XX] Hampshire, Stuart. HIS ESSAY ON SPINOZA AS AN IDENTITY THEORIST. APA?

[Hume, 1980] Hume, David. *Dialogues on Natural Religion.* Indianapolis: Hackett Publishing Company.

[Israel & Perry, 1990] Israel, David and John Perry. What is Information. In *Information, Language and Cognition*, edited by Philip Hanson (Vancouver: University of British Columbia Press, 1990): 1-19.

[Israel & Perry, 1991] Israel, David and John Perry. Information and Architecture. In *Situation Theory and Its Applications, vol. 2*, edited by Jon Barwise, Jean Mark Gawron, Gordon Plotkin and Syun Tutiya (Stanford University: Center for the Study of Language and Information, 1991): 147–160.

[I&P&T, 1993] Israel, David, John Perry and Syun Tutiya. Executions, Motivations and Accomplishments. *The Philosophical Review* (October, 1993): 515–40.

[Jackson, 1986] Jackson, Frank. What Mary Didn't Know. *The Journal of Philosophy* LXXXIII(1986) 291-295.

[Jackson, 1986/97] Jackson, Frank. What Mary Didn't Know. In [B&F&G, 1997]: 567-570. Reprinted from [Jackson, 1986].

[Kripke, 1980/97] Kripke, Saul. The Identity Thesis. In [B&F&G, 1997]: 445-450. This is excerpted from Lecture III of *Naming and Necessity* (Cambridge: Harvard University Press, 1980): 145-155.

[K&A&N, 1997] Künne, Wofgang, Martin Anduschus, and Albert Newen, editors. *Direct Reference, Indexicality and Propositional Attitudes*. Stanford, CA: CSLI-Cambridge University Press, 1997.

[Leibniz, 1714] *Monadology.*

[Levine, 1993/97] Levine, Joseph. On Leaving Out What It's Like. In [B&F&G, 1997]: 543-555.
Reprinted from [D&H, 1993]: 121-136.

[Levine, 19??]

[Lewis, XXXX] Lewis, David. Counterpart theory and quantified modal logic.

[Lewis, 1990/97] Lewis, David. What Experience Teaches. In [B&F&G, 1997]: 579-595. Reprinted from [Lycan, 1990]: 499-519.

[Lewis, 1966] Lewis, David. An Argument for the Identity Theory. *Journal of Philosophy*, LXIII No. 1: 17-25.

[Loar, 1990/97] Loar, Brian. "Phenomenal States," in [B&F&G, 1997], 597-616. An earlier version appeared in*Philosophical Perspectives* 4:81-108.

[Locke, 16??/1975] Locke, John. *Essay on Human Understanding.* Edited by Peter H. Nidditch (Oxford: Oxford University Press, 1975. Chapter on Personal Identity.

[Lycan, 1990] Lycan, William G. *Mind and Cognition.* (Oxford: Blackwell, 1990).

[Lycan, 1990] Lycan, William G. A limited defence of phenomenal information. In [Metzinger, 1995]: 243-258.

[Mach, 1914] Mach, Ernst. *The analysis of sensations*, translated by C.M. Williams and Sydney Waterlow, (Chicago & London: Open Court, 1914), p. 4n.

[McMullen, 1985] McMullen, Carolyn. "'Knowing What It's Like',". *Philosophical Studies* 48 (1985) 211-233.

[McTaggart, 19??] McTaggart, John McTaggart Ellis. *The Nature of Existence.*

[Metzinger, 1995] Metzinger, Thomas, editor. *Conscious Experience.* Schöningh: Imprint Academic, 1995.

[Nagel, 1974/97] Nagel, Thomas, What Is It Like to Be a Bat? in [B&F&G, 1997]: 519-527. Reprinted from [Nagel, 1974].

[Nagel, 1974] Nagel, Thomas. What Is It Like to Be a Bat? *The Philosophical Review* LXXXIII (1974): 435-50.

[Nagel, 1983] Nagel, Thomas. The Objective Self. In Carl Ginet and Sydney Shoemaker, eds., *Knowledge and Mind*, 1983.

[Nemirow, 1979] Nemirow, Laurence. *Functionalism and the Subjective Quality of Experience*. (Doctoral Dissertation, Stanford Philosophy, 1979).

[Nemirow, 1980] Nemirow, Laurence. Review of Thomas Nagel's *Mortal Questions*. *Philosophical Review*, 89, 1980: 475-76.

[Nemirow, 1989] Nemirow, Laurence. Physicalism and the cognitive role of acquaintance. In W.G. Lycan (ed.) *Mind and Cognition: A Reader*. (Oxford: Basil Blackwell, 1989).

[Nida-Rümelin, 1997] Nida-Rümelin, Martine. The character of color predicates: A phenomenalist view. In [K&A&N, 1997]:

[Nida-Rümelin, 1995] Nida-Rümelin, Martine. What Mary couldn't know: Belief about phenomenal states. In [Metzinger, 1995]: 219-241.

[Perry, 1977] Perry, John. Frege on Demonstratives. *Philosophical Review*, LXXXVI, no.4 (1977): 474–97.Reprinted in [Perry, 1993].

[Perry, 1981] Perry, John. "The Problem of the Essential Indexical." *Nous*, 1981. Reprinted in [Perry, 1993].

[Perry,1986] Perry, John. Thought Without Representation. *Supplementary Proceedings of the Aristotelian Society*, vol. 60 (1986): 263–83. Reprinted in [Perry, 1993].

[Perry, 1990] Perry, John. Self-Notions. *Logos*, 1990: 17-31.

[Perry, 1993] Perry, John. *The Problem of the Essential Indexical*. (New York: Oxford University Press, 1993.)

[Perry, 1997] Perry, John. Rip Van Winkle and Other Characters. European Review of Philosophy, Volume 2, 1997: 13-40.

[Perry, 1997a] Indexicals and Demonstratives. In Robert Hale and Crispin Wright, eds., *Companion to the Philosophy of Language*, Oxford: Blackwells Publishers Inc., 1997.

[Perry, 1997b] Reflexivity, Indexicality and Names. In [K&A&N, 1997].

[Perry, forthcoming] *Reference and Reflexivity*. (Stanford: CSLI Publications: forthcoming).

[Perry & Israel, 1991] Perry, John and David Israel. Fodor on Psychological Explanations. In *Meaning in Mind*, edited by Barry Loewer and Georges Rey. Oxford: Basil Blackwell, 1991, 165–180 Reprinted in [Perry, 1993].

[Place, 196?] Place, U.T. Place's article.

[Shoemaker, 1997] Shoemaker, Sydney. The Inverted Spectrum Argument or something like that, in [B&F&G, 1997].

[Shoemaker, 1970] Shoemaker, Sydney. Persons and Their Pasts. *American Philosophical Quarterly*, October, 1970.

[Shoemkaer, 1984] Shoemaker, Sydney. *Identity, Cause and Mind*. Cambridge: Cambridge University Press, 1984.

[Shoemaker, 1996] Shoemaker, Sydney. *The First-Person Perspective and Other Essays*. Cambridge: Cambridge University Press, 1996.

[Smart, 1959] Smart, J.J.C. Sensations and Brain Processes. *Philosophical Review* Vol. 68, 1959:141-56.

[Villanueva, 1995] Villanueva, E. *Philosophical Issues*. (Atascadero: Ridgeview, 1995).

[Wettstein, 1981] Wettstein, Howard. Demonstrative Reference and Definite Descriptions. *Philosophical Studies* 40: 241-57.