Union College
Union | Digital Works

Honors Theses | Student Work

6-2015

Our Consciousness as Evidence for Modal Views in the Philosophy of Mind

Daniel Pallies
Union College - Schenectady, NY

Follow this and additional works at: https://digitalworks.union.edu/theses

 Part of the Epistemology Commons

Recommended Citation
Pallies, Daniel, "Our Consciousness as Evidence for Modal Views in the Philosophy of Mind" (2015). *Honors Theses*. 370.
https://digitalworks.union.edu/theses/370

This Open Access is brought to you for free and open access by the Student Work at Union | Digital Works. It has been accepted for inclusion in Honors Theses by an authorized administrator of Union | Digital Works. For more information, please contact digitalworks@union.edu.

Running Title: Self-Sampling and Consciousness

Our Consciousness as Evidence for Modal Views in the Philosophy of Mind

By

Daniel Pallies

* * * * * * * * *

Submitted in partial fulfillment

of the requirements for

Honors in the Department of Philosophy

UNION COLLEGE

March, 2015

ABSTRACT

PALLIES, DANIEL Our Consciousness as Evidence for Modal Views in the Philosophy of Mind. Department of Philosophy, March 2015.

ADVISOR: David Barnett

In this paper I argue that we have evidence to believe certain views in the philosophy of mind over others. Specifically, the fact that humans are conscious is evidence in favor of a view insofar as that view holds that a greater proportion of beings in the universe are conscious. My reasoning is as follows: if a greater proportion of beings in the universe are conscious, then it is more likely that human beings will be conscious; human beings are conscious, therefore we have reason to believe that a greater proportion of physical beings in the universe are conscious. Human consciousness is evidence in favor of views which hold that a greater proportion of beings are conscious, because human consciousness is *more likely,* epistemically speaking, under those views. Once I have made this argument, I then turn to familiar views in the philosophy of mind. I argue that under certain metaphysical views about the relationship of the physical to the mental, we ought to think that a greater proportion of beings are conscious. Therefore human consciousness is evidence in favor of those metaphysical views.

Chapter 1: Facing Up to the Existence of Consciousness

In this paper I will explore a novel method for determining the relationship between the physical and the mental. Specifically, I will argue that our existence as conscious beings is evidence which can help us decide which theory about the relationship of the physical to the mental is correct. My project will be broken up into a few distinct steps. First, I will defend the concept of the mental. I will argue that there is such a thing as the mind, consciousness, or conscious experience. Second, I will argue that a physical analysis of the brain cannot help us decide which physical-mental identity theory is correct. There is such a thing as consciousness, and we cannot determine how it arises from the physical merely by analyzing the physical. Third, I will introduce my own argument in favor of certain physical-mental identity theories. My argument is not based in physical analysis; rather, it is an epistemological argument according to which our own existence as conscious beings is a kind of evidence. I call my argument the *self-sampling argument.*

1.0 Introduction to the Mental

What is the mental? As I use the term, I refer to the experience one has as a living, thinking, and perceiving being. For example, I have a certain experience as a type these words: I feel the keys under my fingers, see a screen in front of me, taste lukewarm coffee, and so on. These sorts of experiences comprise what we often call "the mind," or consciousness. Conscious experiences are far from being mysterious to us—in fact, they may be the things of which we are most intimately aware.

However, it would be a mistake to think that most people share common intuitions about consciousness; from my own experience, it is obvious that different

people approach the subject of the mental in very different ways. Some people will consider the above description to be a very intuitive and even tediously obvious description of conscious experience. On the other hand, some others will think that it fails to answer the question or is perhaps even incoherent. One of the most challenging things about the philosophy of mind is that everyone arrives at the central problems with their own preformed intuitions. What seems intuitive to one person will seem ridiculous to another, for reasons that can often be very difficult to articulate.

In the proceeding section I will do my best to articulate some of these intuitions. In order to have any kind of discussion about the mental and its relationship to the physical, we must be on the same page with the meaning of "mental." My hope is that we can arrive at a shared understanding of what I mean by "the mental" before moving on to the rest of the paper. If you are already on board with my above description of the mental, you can safely skip this section.

1.1 Intuitions about the Mental

When I describe the mind as conscious experience, it is sometimes suggested that my description leaves something out: namely, what kind of *thing* the mind is. Without knowing what kind of thing consciousness is, it is suggested, we do not really know anything about it at all. Is the mind something we can touch? Or is it something ethereal and mystical, perhaps a soul? I suspect that many people worry that any discussion of the mind will involve the existence of some mystical or mysterious mental substance. But as far as I am concerned, my description of the mind as mental experience is actually agnostic with respect to these questions. However the mind exists, whether it has its basis

in the physical brain or an immaterial soul, it can still be accurately described as conscious experience.

Let us posit that the mind has its basis entirely in the brain—the mind *is* the brain, in some important sense. Even so, we can still talk about our conscious experience without explicitly referring to the brain itself, *and* without positing that conscious experience is literally a physical thing. An analogy is helpful here. Consider the way that we refer to things in a game of poker. Obviously the objects involved in a poker game are wholly physical things, but can refer to the object and events in a poker game with words like "bets," "flushes," "flops," and so on—none of which sound like terms for strictly *physical* things. What is the physical composition of a "*bet*?" To use such terms does not mean that something non-physical is going on in a game of poker. Just as a "bet" has its basis in the physical world without itself being a literally physical thing, conscious experiences can have their basis in the physical world without being literally physical things.

Consider a paradigmatic example of a conscious experience. When I look at a red sheet, I say I have an experience of the color red. "An experience of the color red" does not sound like a physical thing, but then, neither does "bet." Just as one can think that nothing non-physical is going on in a poker game and still use the word "bet," one can think that the process of vision is entirely physical and still use the description of "an experience of the color red." To say that the redness of the sheet has a felt quality for me—that I see and experience redness—is not to say that there is something non-physical going on. One could maintain that one has an experience of red, and also that the

experience has its basis a physical process. That is, one could maintain that referring to "an experience of the color red" is a way of speaking about a physical process.

For my purposes, I have no need to commit one way or the other. I am remaining agnostic with respect to whether or not conscious experiences are wholly accounted for by physical states or processes. My point is only that one can use the term "conscious experience" meaningfully, without committing to whether or not conscious experiences can be explained entirely in physical terms. As a consequence, the term "conscious experience" is somewhat strange. If one believes that conscious experiences have their basis entirely in the brain, then by "conscious experience" one refers to a particular phenomenon which has its basis in physical phenomena in the brain. If, on the other hand, on believes that conscious experiences have their basis in something else, then the meaning of "conscious experience" will be different for that reason. Nevertheless, those two speakers' uses of the phrase conscious experience will overlap in an important way: they will both refer to the consciously felt experience, whether this experience has its basis in the brain or elsewhere. This overlap—that part of the meaning of "conscious experience" which refers to the consciously felt experience rather than its basis—is what I mean by "the mental."

Despite my efforts to remain agnostic on controversial issues, it is impossible for me to present an account of the mind which will satisfy everyone. My account of the mental includes some positive theses with which some people will disagree. For example, on my account of consciousness, the contents of consciousness are essentially private phenomena. My experience of the taste of coffee is only *my* experience—only I have immediate access to it. I know about my conscious experiences in a special way that is

not available to others. Other people can *infer* that I am having a particular experience, but they do not are not immediately aware of it in the same way that I am.

In the proceeding section, I will defend this epistemic thesis. We do have special epistemic access to the contents of our consciousness; we know that we are having conscious experiences in a special way. It's certainly possible that other people can draw inferences about our conscious experience based on our behavior or some neurological data, but our own access to our conscious experience is non-inferential. When we have a conscious experience of something, we know about that conscious experience just by virtue of having it. I will refer to this view as the *special access view*.

2.0 Against the Special Access View

Daniel Dennett is one philosopher who believes that the privileged access view is wrong. In this section I will defend my view against a particular argument formulated by Dennett. I will focus on sensory perceptions because they are the most popular example of the contents of consciousness. "Sensory perceptions" are just what I have described earlier: things like itchiness, color, taste, and so on. I maintain that sensory perceptions are both *private* and *ineffable*. Roughly speaking, this means that they cannot be described to someone who has not experienced them, and that they can't be known by anyone except the person who apprehends them. I can't explain to a blind man what colors look like, for example, nor can I gain access to another person's mind and learn what she sees, feels, perceives, and so forth. If sensory perceptions are private and ineffable, then we do have privileged access to them. Others cannot know about them in the way that we do, and we are incapable of sharing our knowledge of them. Dennett denies that sensory perceptions have these properties as I understand them.

To overturn the special access view, Dennett invokes a variant of the familiar *inverted spectrum* thought experiment. The thought experiment is supposed to demonstrate that we do not really have any sort of privileged access to our mental states. We do not know about our sensory experiences in a special way. The thought experiment is as follows:

> Imagine that as you sleep one night, an evil neurosurgeon conducts a sinister procedure on your brain. You wake up the next day with no memories of the neuroscientist or the procedure, but it is nevertheless obvious to you that something has happened. As a result of the neurosurgery, the colors of the world around you seem all wrong. In fact, they have been reversed. What you once called "red" you now see as "green;" "yellow" is "purple," and so on. The experience is very disorienting, and it only gets worse as you realize that the situation may be more complicated then you'd first imagined. Perhaps the evil neurosurgeon didn't change your perceptions of color after all, but rather your memory impressions concerning colors[1]. After all, if your experience of color is not beyond the grasp of this advanced neurosurgery, why should you assume that your memories are safe?

The victim's conundrum takes the following form. She looks up at the sky and assigns the word "orange" to the color she sees. It is possible that before the sinister procedure, the sky really did look "blue" to her, as she uses the word "blue" now. But

[1] Dennett describes the neurosurgical process as altering either your color experience, or your memory-linked color experience reactions. I do not know what a change in one's memory-linked color experience reactions might be, except a change in one's memory of one's color experiences. For the sake of easy reading, then, I express the second kind of neurological change as a change in one's memory of color experiences.

then again, maybe it has always looked the same to her. Perhaps she only calls the sky "orange" now because her memories have been altered such that she seems to remember the sky being a different color before—what she calls "blue" now. In that case, only the way in which she assigns words to various sense impressions would be changed. Her actual sensory impressions would have remained static, despite the fact that she really are is dismay about the current state of her color experience.

To make matters worse, it seems like no future experience on the part of the neuroscience victim will help her out of her conundrum. She still will not know what colors she is seeing. The question is: are they the same colors she saw before, or have they been switched?

In more traditional formulations of the inverted spectrum thought experiment, the hapless victim was not quite so profoundly flummoxed by the state of her visual experience. Traditionally, the victim is at least aware of the change in conscious experience. Others, of course, would not know that there was any change in the victim's color experience. But as Dennett's thought experiment demonstrates, even the victim herself can fail to know the relevant facts about her own color experience. She might not know whether or not her color experience had changed at all.

If the victim does not know whether or not her color experience has changed, then it seems she has no better access to her color experience than does anyone else. But if it can be possible for one to not know about the status of one's own sensory experience, it seems like such experiences are not directly accessible after all. In this case, the hapless victim's best bet would be to track down the neuroscientist and have her explain which procedure she performed. Far from having "special access" to her own mental life, the

victim would be better off relying upon someone else's testimony to determine the contents of her own visual experience! The neurosurgeon would be in a better position to know the contents of her conscious experience than she herself would be. The thought experiment seems to disprove the special access view, according to which a person has more direct and immediate access to her own conscious states than do other people.

2.1 In Defense of the Special Access View

Dennett's thought experiment is intended to overturn the view that sensory states are *internal* in the sense that is usually ascribed to them. If someone else knows better than I do about my own sensory states, than it certainly does not seem like my sensory states are internal. It also seems evident that I do not have special access to them.

I think Dennett makes an important mistake. His mistake is in misunderstanding the sense in which color experiences are "internal." Dennett explains the thought experiment as though color experiences are supposed to be internal to individual *people,* whereas it is more appropriate to say that color experiences are internal to individual *experiences of consciousness.* Usually, there is no problem in using these two terms interchangeably, since it seems as though any given person has a single experience of consciousness. However, as Dennett's thought experiment shows, this is not necessarily the case.

In an important sense, it seems that the victims of the evil neurosurgeon have had two separate and distinct experiences of consciousness. After all, their previous color experiences are not within the scope of their current experience. That is to say, they cannot be summoned up in the victims' current mental states, and so information about the victims' color experiences pre-surgery is external to those same victims post-surgery.

The entire continuum of color experiences is internal to the victims in a sense, because the experiences can always be attributed to the same *person*. It is the same person who had access to his color experiences one day and did not have access to them on the next day. This, however, does not seem to be the important sense in which perceptual experience is supposed to be internal.

In fact, the victim seems to be ignorant about her previous color experiences in the same way that we are ignorant about the perceptual experiences of other people in general. Her previous color experiences are no more a part of her current mental state than are the mental states of other people. Therefore, it does not seem that her previous mental states really are internal to her in any meaningful sense.

Of course, if the neurosurgeon only altered the victim's *memories* of color experience, than she really does have authentic access to her memories of color experience. What matters, however, is that she is in doubt about the authenticity of those memories. The thought experiment tells us that she has no way of knowing whether or not her memories actually represent her past mental state. What she "seems to remember" will not help her, because she has no reason to think that her memories are accurate. She is ignorant as to the accuracy of her memories whether or not those memories have actually been altered.

If the victim's memories are not a part of her current conscious experience, then Dennett's argument seems to lose its force. To see why, imagine that you are the victim's confidant. She asks you whether or not her color experiences have been inverted—whether or not the sky appears blue or orange to her, for example—and of course you have idea. You are in total ignorance as to whether or not *her* color experiences have

changed. This conventional sort of ignorance is the only sort of ignorance the victim herself experiences if her pre-surgery color experiences are entirely external to her post-surgery.

Dennett's thought experiment does not seriously endanger the special access view—we just need to clarify *who exactly* has special access. Dennett takes the prevailing view to be that a *person* has special access to her sensory experiences. He successfully demonstrates that an individual *person* does not always have special access to her own mental states. That is fine by me. I think it is more accurate to say that an *individual experience of consciousness* has special access to her sensory experiences. If there was no epistemic break in the victim's conscious experience which caused her to doubt her past experience, she would be able to determine the present state of her experience. That is, if she knew that her memories had not been altered, she could simply compare her present experience with her past experience to determine if there was any difference. But there was an epistemic break in the victim's conscious experience, so for all intents and purposes the victim had two distinct experiences of consciousness. Usually a person just is an individual experience of consciousness; but, as Dennett demonstrates, this is not always the case. In those normal cases in which we have just one experience of consciousness, we do have special access to our conscious states. We directly know about what we are consciously experiencing, in a way that others do not.

Chapter 2: Locating Consciousness

1 Introduction

We should take the existence of consciousness seriously. There really is something it is like to have conscious experiences of color, taste, texture, and so on.

However, accepting that consciousness exists will lead us to take on the problem of explaining its existence, and this is a difficult problem indeed.

First of all, it is difficult to imagine how consciousness could possibly have its basis in the physical world. At this moment I see a brown coffee stain on the desk in front of me, taste peppermint gum in my mouth, and feel the texture of the keyboard under my fingers. Now we have some vague sense that the neuronal action of my brain is responsible for these sensations—but how? How could it be that the action of my sense organs and nervous system makes these things look, taste, and feel the way they do to me at this moment? How could mental experience have its basis in a physical system, however complicated? This might be called the *metaphysical mind-body problem.*

However, there is another obstacle in establishing the relationship of consciousness to the brain. Let us take for granted that the brain is the physical basis of consciousness. The question remains: how could we come to know how brain processes are responsible for mental states? There are a plethora of theories which take the brain to be the physical basis of consciousness, but these theories still differ importantly from one another. One such theory holds that mental states are identical with brain states. Another holds that mental states are identical with functional states. Without going into detail about the vicissitudes of these theories, we can at least say that they differ with respect to what they take to be the relationship between the physical and the mental. But they both are with consistent with the view that the brain is the seat of consciousness. How then can we determine which view is correct? Let us say that certain neuronal states or configurations are the correlates of certain conscious states—how would we know which

such neuronal states match up with particular conscious states? This I call the *epistemic mind-body problem.*

One way in which we might hope to determine which theory is true is by analyzing physiological processes. We take the brain to be the physical basis for consciousness. If we study the physical side of the physical-mental relationship, we might determine the nature of the relationship to the mental. If the relationship of the physical to the mental is such that conscious experience is identical with some physical feature of the brain, then it might be possible to physically locate it. Perhaps if we had an extremely advanced knowledge of the brain, we would understand how some physical phenomena constitute or are responsible for consciousness. We would be able to demonstrate that a particular physical phenomenon is identical with a particular mental phenomenon. By repeating this process we could arrive at a complete map of the human brain and mind.

This method of scientific inquiry has succeeded in explaining other natural phenomena. Take the phenomena of heat, for example. Before one knows certain scientific facts in chemistry and physics, it is difficult to imagine how heat could be the same thing as molecular motion. However, once we know how molecules operate on a small scale, we can see how that phenomenon constitutes the phenomenon of heat. Perhaps the same sort of analysis will explain consciousness. Once we have a sufficiently complex understanding of the physical brain, we will see how its properties naturally provide the basis for consciousness. We will intuitively understand that consciousness *just is* the action of the brain in the same way that heat *just is* the motion of molecules.

In this chapter I will argue that we should be pessimistic about the effectiveness of this approach. It does not seem possible that the same kind of reductive analysis which

works for the concept of liquidity will also work for the concept of consciousness. We can understand liquidity by understanding the physical particles which are the basis of liquidity, but we cannot understand consciousness simply by understanding the physical matter which comprises the brain. We can determine analytically that liquidity is constituted by a certain kind of motion of molecules, but we cannot determine analytically that consciousness is some kind of physical phenomenon. For any theory T which holds that some physical phenomenon constitutes consciousness, we cannot prove that T is true simply by analyzing the physical phenomenon in question.

2 Classic Arguments

In his *Naming and Necessity,* Saul Kripke famously argues that there are serious challenges facing philosophers who want to identify particular physical phenomena with particular mental phenomena. I restate his argument here in order to motivate the discussion that follows.

Some names of things are **rigid designators**. Rigid designators refer to the same thing across all possible worlds. A possible world is a stipulated world in which certain contingent facts about our own world are different. Consider the person "Saul Kripke." Saul Kripke might have done many things which he did not actually do in the actual world: he might never have given the *Naming and Necessity* lectures, he might have died in a plane crash, or he might have never even been born. That Saul Kripke was born, lived up to the present, and did give the *Naming and Necessity* lectures are all contingent truths about our world. "The person who gave the *Naming and Necessity* lectures" is not a rigid designator, because it might refer to different people in different worlds. Someone else might have given those lectures. What is not contingently true is that Saul Kripke is

Saul Kripke. In every possible world, Saul Kripke is Saul Kripke. Therefore, "Saul Kripke" is a rigid designator.

Imagine that the person we call "Saul Kripke" was never born, but that instead someone else grew up to do everything that Saul Kripke did in life. He gave the *Naming and Necessity* lectures, became a professor at Princeton, and so on. That person would not be Saul Kripke, because the name "Saul Kripke" does not refer to whatever person performed certain actions and life and fit a certain description. It refers *rigidly* to a particular person: the Saul Kripke of our world. Saul Kripke might have lived an entirely different kind of life, but he would still be Saul Kripke. And someone else might have lived Saul Kripke's life, but that would not make him Saul Kripke. This is what is meant by the statement that "Saul Kripke" is a rigid designator.

Proper names are not the only rigid designators. "Natural kinds" are another kind of thing which are rigidly designated by the names we assign to them. A natural kind is a unified category of things found in nature, such as "tigers," "quarks," or "gold." Consider the way these names pick out classes of objects or phenomena in our world. The word "gold," for example, refers to a *particular kind of thing found* in our world; it does *not* refer to a particular description of a thing. The name points at a particular substance: that stuff which we call "gold." This may seem obvious, but consider the alternative. Let us say that we use the word gold to mean "a lustrous yellow metal." We can imagine a world other than our own in which a metal with that description exists; but for all of its similarities to gold, that substance might or might not turn out to actually be that substance we call "gold" in our world. Despite all its superficial similarities to gold, it might turn out to have an entirely different chemical composition. If we were to find out

that this lustrous yellow metal in this other world was actually comprised of spooky ectoplasm, we surely would not conclude "the gold in this world is composed of spooky ectoplasm." We would say that this substance was something else entirely, something which superficially resembles gold. The word "gold" refers to a particular category of stuff in our world, not to any kind of stuff which meets a particular description.

Kripke claims that rigid designators have important ramifications for the philosophy of mind. Consider the view that a particular physical phenomenon, such as C-fiber stimulation, is identical with a particular mental phenomenon, such as pain. Someone who takes this view will claim that C-fiber stimulation is pain. Saul Kripke argues that this claim is dubious. When we say that the natural kind of C-fiber stimulation is identical with the natural kind of pain, both "C-fiber stimulation" and "pain" are rigid designators. The stimulation of my C-fibers refers to a particular phenomenon; across all possible worlds, the name "C-fiber stimulation" refers to the same kind of thing. Similarly, "pain" refers to a particular natural kind. Kripke claims that if two rigidly designated phenomena are identical, then it must be a necessary truth that they are identical. But it is not easy to see how one could explain that the connection between physical states and mental states is true *as a matter of necessity*. Therefore, according to Kripke, there is a serious problem for philosophers who hold physical-mental identity claims.[2]

The upshot of Kripke's argument is this: in order to demonstrate that a certain physical phenomenon constitutes a certain mental phenomenon, one must demonstrate that the one phenomenon *necessarily* constitutes the other. One cannot claim that C-fiber firing can sometimes constitute pain any more than one can claim that a certain motion of

[2] Kripke, Saul (1980) 'Naming and Necessity' 144–145.

molecules can sometimes constitute liquidity. If the motion of molecules is of a certain sort, then that molecular motion *just is* liquidity. In the same way, one must demonstrate that if one's C-fibers are firing, those firings *just are* pain.

No part of the preceding argument shows that it is impossible for C-fiber firings to be pain. It merely sets up the problem for materialist philosophers who argue that such identities exist. The materialist philosopher would have to show that C-fiber firings are pain in the same way that liquidity is a certain kind of molecular motion. As the proceeding sections will demonstrate, this is a very tall order.

2.1 The Knowledge Argument

Perhaps the best known thought experiment in the philosophy of mind is employed by Frank Jackson in his knowledge argument. The thought experiment is as follows. Mary is a brilliant neurosurgeon who is born and raised in a black and white room. There she learns everything there is to learn about the neuroscience of vision. She learns about the physics of light and how particular wavelengths interact with particular cone cells in the retina, and so on and so forth. She learns all the physical facts about color vision. However, she does not actually see any colors. One day she is freed from the room and is exposed to the colors of the outside world. Frank Jackson tells us that at this point Mary learns a *new* fact about color vision: she learns what it is like to see colors.[3]

It certainly seems plausible that Mary learns something new by seeing colors for the first time. Before leaving the room, she had no idea what it was like to see the color red. She could not have summoned up an image of the color red had she been asked to do

[3] Jackson, Frank (1982) 'Epiphenomenal Qualia' *Philosophical Quarterly* 32:127-135.

so. But after seeing the color red, she can. This is despite the fact that she knew all the physical facts before leaving the room.

Frank Jackson's knowledge argument is constructed around this thought experiment. He argues:

(1) Mary knows all the physical facts.

(2) Mary does not know all the facts.

_____

(3) The physical facts do not exhaust all the facts.

This argument is usually taken as an argument against **physicalism**. Physicalism is the doctrine that everything in the universe is physical. If everything in the universe is physical, then it seems that the only kinds of facts are physical facts. But when Mary saw color for the first time, she learned a non-physical fact. Therefore physicalism is false.

The strategy for toppling physicalism is as follows. The argument begins by establishing an *epistemological* gap between the physical facts and the mental facts. Mary does know all the physical facts, but she does not know all the mental facts. Therefore one cannot deduce the mental facts from the physical facts. Having established that there is an epistemological gap, the argument goes onto say that there is an *ontological* gap between the physical facts and the mental facts. If you cannot deduce the mental facts from the physical facts, then there is a difference between the physical facts and the mental facts. And so:

(1*) There is an epistemic gap between the physical facts and the mental facts.

(2*) If there is an epistemic gap between the physical facts and the mental facts, then the physical facts are not the same as the mental facts.

(3*)There are facts about the universe which are not physical facts.

_____

(4*)Physicalism is false.[4]

For my own purposes there is no need to endorse this controversial argument. I only need to endorse (1*). There is an epistemic gap between the physical facts and the mental facts; one does not understand what it is like to see the color red *merely* from knowing all the physical facts about color vision. One does not understand what it is like to be in pain *merely* from knowing all the physical facts about C-fiber stimulation. And so on. It is not possible to derive facts about the mental merely from analyzing facts about the physical, and so we should be pessimistic about finding physical-mental identities simply by studying the brain.

3 The Shape of the Problem

Identities between physical and mental phenomena would have to be *necessary* identities. It would have to be the case that a certain physical phenomenon *just is* a certain mental phenomenon. In other cases of necessary identities, it is possible to determine the identity by analyzing the identical phenomena. We can determine the identity between liquidity and molecular motion by analyzing the phenomena in question. However, the identities between physical phenomena and mental phenomena do not seem vulnerable to this kind of analysis. The knowledge argument demonstrates that merely

[4] Chalmers, David 2002. 'Consciousness and its Place in Nature' in *Blackwell Guide to the Philosophy of Mind,* S. Stich and T. Warfield eds., Blackwell.

studying the physical facts will not tell us about their relationship to the mental. So it is not clear how we could discover the physical-mental identities.

It is important again to address the difference between the metaphysical and epistemological problems we face. The metaphysical problem is that we do not know how physical phenomena could constitute mental phenomena. One might invoke the knowledge argument in order to demonstrate this problem. One could claim that the physical brain cannot be identical with mental phenomena, because Mary knows all the physical facts about the brain but does not know about the mental phenomena. If a certain physical action of brain *just is* the sensation of the color red, then Mary should know about the color red by virtue of knowing about the relevant physical action of the brain. But she does not know about the color red. Therefore the physical phenomenon in the brain is not identical with the mental phenomena of color perception.

This is not the lesson I am drawing from the knowledge argument. I am instead invoking the epistemological problem. Let us grant that the metaphysical problem is solvable; somehow it turns out that certain physical processes in the brain are identical with certain conscious experiences. So, for example, physical phenomenon P1 *just is* that mental phenomenon M1. We still face a problem. Even if it is possible for P1 to be identical with M1, it is unclear how could we come to know that P1 is M1. Even Mary's complete knowledge of the brain did not demonstrate that P1 is identical with M1.

If we cannot determine which physical phenomena are identical with which mental phenomena, then we will face a problem in trying to decide between competing theories which differ with respect to how they describe the relationship between the physical and the mental. For any theory T which holds that a given mental phenomenon

is identical with a given physical phenomenon, we would need to learn that the relevant mental-physical identity is true in order to know that T is true. But it seems we cannot learn about mental-physical identities just from studying the physical.

Suppose that there are two competing theories about the relationship of the physical to the mental, T1 and T2. According to T1, the conscious experience of red R is identical with a certain physical state of the brain P. According to T2, R is identical with a certain functional brain state F. Let us also say that our best neuroscience cannot rule out either of these theories. When human beings experience red, they also are both in the physical state P and the functional state R. We therefore cannot determine the truth of either T1 or T2 on an empirical basis. And we cannot determine the truth of T1 or T2 by analyzing the concepts involved—nothing about P or F lets us see how it could be identical with R.

The differences between T1 and T2 could very well be substantive differences. One can imagine an advanced robotic brain which succeeds in actualizing physical state P but fails to actualize functional state R. On T1, that robot could have an experience of the color red (and would necessarily have that experience as long as state P obtains). On T2, it could not have that experience. An analysis of P and F, no matter how thorough, will not help us understand how those phenomena could be R. An empirical approach seems no more promising. We could ask the robot what it sees, but we would have no way of verifying whether or not that report is accurate. Siri claims to have various mental processes, but (presumably) she has no real mental life. We therefore need a different kind of evidence to decide between theories like T1 and T2.

3.1 Knowledge about Physical-Mental Identities

The knowledge argument seems to demonstrate that one could not understand physical-mental identities merely by studying the physical. But it does not tell us why we could not come to know those identities. It implies that there is some big difference between physical-mental identities and other kinds of necessary identities, like the liquidity-molecular motion identity. It seems Mary could understand liquidity completely if she had complete physical knowledge of liquid water. But she could not know about the sensation of red even with complete physical knowledge of the brain. In this section I will attempt to explain the difference between the identities which accounts for this difference in epistemic status. First I will illustrate a paradigmatic example of a necessary identity which can be analyzed and discovered. Then I will contrast this identity with the identity of pain to C-fiber firings. This identity cannot be analyzed or discovered in the normal way.

One example of a necessary identity is the identity of Phosphorous to Hesperus. Both "Phosphorous" and "Hesperus" refer to certain celestial bodies. The names "Phosphorus" and "Hesperus" are rigid designators; they refer to the same things across all possible worlds. Though Phosphorus might have been shunted into the sun or collided with the Earth, it could not have failed to be Phosphorus. The same holds true for the heavenly body of Hesperus. As it turned out, both Phosphorus and Hesperus are names for the same planet: Venus. The heavenly body was called "Phosphorus" when it appeared in the morning sky and "Hesperus" when it appeared in the evening sky, but the two words were being applied to the same thing. Because the name "Phosphorus" points to the same thing as does "Hesperus," Hesperus could not fail to be Phosphorus any more than Phosphorus could fail to be Phosphorus. "Hesperus" and "Phosphorus" always point

to the same thing (if they point to anything at all, after all, it's possible the celestial body might never have formed in the first place) across all possible worlds. Because the identity is true across all possible worlds, the identity is necessary. Phosphorous *just is* Hesperus.

It was not immediately apparent that Phosphorus was identical with Hesperus, because our initial impression of the two heavenly bodies was that one appeared in the sky during the morning, and the other appeared in the sky in the evening. It was these original descriptions which we used to describe the referents of the words "Phosphorus" and "Hesperus," respectively, and these descriptions do not imply that Phosphorus is identical with Hesperus. These descriptions, however, merely provide incidental properties of the heavenly bodies. "Phosphorus" refers to *that thing* which we see in the morning sky, regardless of whether it is in the morning sky or not. When we discovered that Hesperus is Phosphorus, we moved past the incidental descriptions of the two heavenly bodies. We instead looked at the things to which the names "Hesperus" and "Phosphorus" pointed, and found that they pointed at the same thing, the same celestial body.

"The heavenly body which appears in the morning sky" is a non-rigid designator. It applies to Phosphorus, but not as a matter of necessity. The hunk of matter we call "Phosphorus" could appear elsewhere in the sky, or not at all. Likewise, "The heavenly body which appears in the evening sky" is also a non-rigid designator. It is a contingently true description of Hesperus. Let "C_P" and "C_H" stand for these contingently true descriptions of Phosphorus and Hesperus, respectively. C_P is not necessarily identical with C_H. However, C_P is contingently true of Phosphorus, and Phosphorus is necessarily

identical with Hesperus, which in turn is described in a contingently true way by C_P. So although C_P and C_H do not *appear* to be identical, the rigidly designated subjects they describe *are* identical. And because the subjects they describe are rigidly designated, they are identical as a matter of necessity.

This is the normal process whereby we explain how two things are identical as a matter of necessity. We look for the rigidly designated subjects picked out by the contingent descriptions of the two things in question. The descriptions could apply to two different subjects, in which case the two things are not identical. For example, it could have turned out that the planet Venus never existed at all, but that instead there were two distinct heavenly bodies in the sky, one in the morning and one in the evening. In that case, there would have been no identity whatsoever between those two heavenly bodies picked out by the contingent descriptions of those bodies. "The heavenly body which appears in the morning sky" and "the heavenly body which appears in the evening sky" would not be identical in that possible world— and they would not be the heavenly body called "Hesperus" and "Phosphorus" in our world.

On the other hand, if the rigidly designated subjects are the same—as they are in our world—then their identity with one another is a necessary truth. In our own world, the perspective-specific description of Phosphorus picks out the same referent as the perspective-specific description of Hesperus, even though the two perspective-specific descriptions are themselves different. Our empirical discovery was consisted of overcoming the limited perspectives which had led us to think that Hesperus and Phosphorus were different things.

Now we can return to the case of C-fiber firings and pain. Here, the normal method meets difficulty. As in the normal cases, we begin with two phenomena which are not immediately or intuitively identical with one another: namely, the C-fiber stimulation and the pain. Normally, this seeming difference could be explained by showing that while the contingent descriptions we associate with the two phenomena are different, the phenomena themselves are not. We could show how "C-fiber stimulation" and "pain" both rigidly designate the same thing, just from different perspectives. But as regards pain, the perspective seems to be built in, such that there is no perspective-free account of what a pain is. In the case of Hesperus and Phosphorus, the contingent descriptions were descriptions of how the phenomena in question *seem* to us. When we describe Phosphorus as "the heavenly body which appears in the morning sky," we are providing a description of how Phosphorus looks from a certain perspective. In order to show that Phosphorus was identical with Hesperus, we separated this perspective-based description from the rigid designator of "Phosphorus" itself, and showed that the rigid designator points to the same thing as does "Hesperus." When we talk about pain, however, we only ever talk about how pain feels or seems to us. There is no distinction between the perspective-based description of pain, and the rigid designator "pain." "Pain," the rigid designator, can only point to the way pain feels to a given person, because pain *just is* a feeling.

Put another way, any description of pain seems like it must have a first-person ontology, because pain is a first-person phenomenon. When we described the way Hesperus and Phosphorus looked to us, we were describing first-person phenomena in that case, too. We were describing how the heavenly body appeared *to an observer*. It

appeared as a heavenly body in the evening sky and in the morning sky, respectively. As regards Hesperus and Phosphorus, it was possible to conceive of a third perspective which would encompass the previous two, and reveal their limitations. We can imagine having a God's eye view of the planet Venus, for example, in which case it would be obvious to us that Hesperus and Phosphorus are the same planet, and equally obvious why people on Earth had been wrong in thinking that they weren't. Hesperus and Phosphorus are obviously identical to one another from that wider perspective, even though certain first-person perspectives make it seem as though they are not. Contrastingly, it seems impossible for there to be any analogous perspective which would clearly demonstrate the identity between C-fiber firings and pain. Pain, it seems, only permits of being seen from one perspective—the perspective of the person who experiences it.

There just is not any perspective we can take to see how pain is the same thing as C-fiber firings, because pain is inherently perspectival. Perhaps beings other than ourselves have the imaginative or cognitive capacities to understand how C-fiber firings are a priori identical with pain, but we do not have those capacities. For us, there is no a priori link from the physical to the mental.

We conceive of the physical and the mental in very different ways. Physical things are spatially extended—they exist in space and have spatially defined properties. Mental things are not spatially extended; my pains do not occupy or fill space; they have no concrete location. Mental things have a raw feel which can be explored through introspection. There is no raw feel of a chair in the same way that there is a raw feel of pain. The properties of physical phenomena are of no help in explaining the mental

phenomena, and vice versa. Mental phenomena cannot be defined in spatial terms, so our knowledge of the spatially defined properties are of no help to us in understanding mental phenomena. And the physical substance of our brain is not the sort of thing which has a raw feel, so introspection alone will not help us understand it. Our modes of understanding the mental and the physical simply do not relate to one another.[5]

Some theorists argue that is simply impossible for us to understand how the physical could relate to the mental. I will not take that strong stance here. It may be possible for us to fully grasp the link between physical and the mental by way of some heretofore unforeseen method. I have my doubts, but I admit it is at least possible. I am only laying out the challenge faced by those who want to demonstrate that something like C-fiber firings just are the same thing as the experience of pain.

In the proceeding sections I will examine a number of views which attempt to close the gap between the brain and the mind. I divide these views into two categories. The first category contains those views which might make progress on the metaphysical gap but do not resolve the epistemic gap. In other words, they explain how certain physical phenomena could be certain mental phenomena, but they do not explain how we could know *which* physical phenomena are identical with *which* mental phenomena. They therefore will not help us decide between theories with differ with respect to which physical phenomena they take to be identical with mental phenomena.

The second category is a more serious challenge to my view that there is an epistemic gap. Ned Block and Robert Stalnaker argue that there is no a priori analysis which explains normal macrophysical truths in terms of normal microphysical truths to. So, for example, knowing all the microphysical facts about H20 will not tell us that water

[5] McGinn Citation

is H20. But if a priori analysis does not explain how macrophysical truths arise from microphysical truths, then we should not expect an a priori analysis of microphysical truths to explain how mental or phenomenal truths arise. In that case, the lack of an a priori analysis from the physical to the mental does not indicate that there is a robust epistemic gap between the physical and the mental. Block and Stalnaker's argument is technical and difficult, and it will take some involved analysis to demonstrate where they go wrong. There *is* a priori analysis from microphysical truths to macrophysical truths; there is *not* a priori analysis from physical truths to mental truths.

3.2 Views about Physical-Mental Identities

There are many views which attempt to explain how certain physical phenomena could be identical with some mental phenomena. I examine only a few here; I take these views to emblematic of broader contemporary philosophical projects. These explanations might help us understand how physical things could turn out to be the same as mental things, but they still do not help us close the epistemic gap.

I first examine Brian Loar's theory of phenomenal concepts as recognitional concepts. A recognitional concept, Loar tells us, is a concept we have of something on account of recognizing it. They are concepts which take the form of "that thing is one of that kind." Importantly, they are also perspective-relative. Suppose I see a crow up close and form a recognitional concept of it—I recognize that bird as one of *that kind* of recognizable things. Later, when I see a crow flying far in this distance, I might form a new recognitional concept. Recognitional concepts are rooted in our own experience of the world; they do not refer objectively to objects outside of our experience. So the

recognitional concepts of "that bird (up close)" and "that bird (flying far away)" are a priori distinct, despite the fact that they refer to the same thing in the world.

To say that phenomenal concepts are recognitional concepts is to say that they are perspectival concepts of certain physical things.[6] So "pain" for example, may be a perspectival concept we have towards our own C-fiber firings. Just as "that bird (up close)" and "that bird (flying far away)" ultimately refer to the same thing in the world, "that phenomenon (experienced as pain)" and "that phenomenon (observed as C-fiber firings)" also end up referring to the same phenomenon. Unlike birds, the experience of pain does not permit of being observed from different perspectives—it only admits of a first person, introspective perspective. However, the experience of pain can still refer to something in the world. Just as the recognitional concept "that bird" refers to a particular physical thing, so too can the recognitional concept "this pain I feel" refer to a particular physical thing. Perhaps it refers to the firings of C-fibers.

Let us grant that phenomenal concepts are indeed recognitional concepts, and that they pick out physiological processes in the brain. The question remains: which phenomenal concepts are recognitional concepts of which physiological processes or states? Introspection alone will not tell me whether or not my pain refers to a particular physical phenomenon happening in my body—at most I would get the sense that something (physical) is wrong, but I certainly would not get a feeling that my C-fibers are firing. And for familiar reasons, merely examining the physical brain and body will not tell me which physiological phenomena correspond to mental phenomena. Loar's theory therefore will not help us to distinguish between competing theories about the relationship of the physical to the mental.

[6] Loar, Brian (1990) 'Phenomenal States' *Action Theory and the Philosophy of Mind* 81-90.

In his *Two Conceptions of the Physical,* David Stoljar takes a swing at the same problem. Like Loar's theory of phenomenal concepts as recognitional concepts, Stoljar presents a plausible solution to the metaphysical problem but fails to offer guidance on the epistemic problem. In his view, there are two different kinds of physical properties: those properties which are explained by physical theory, and those intrinsic properties which provide the categorical basis for the aforementioned properties. Stoljar calls these *t-physical* and *o-physical* properties, respectively. O-physical properties ultimately provide the basis for conscious experience.

Stoljar's theory is compelling, but it does not solve the epistemic problem. It will not help us decide between physicalism and functionalism, for example. Are o-physical properties such that they provide the basis for mental states when configured in a particular way? Or would the o-physical properties give rise to conscious experience in a variety of different particular configurations, so long as those configurations play a particular functional role in the behavior of a system? Merely noting the correlations between brain states and mental states will not help us decide, for familiar reasons. At any given time and for any given brain, both a physical state and a functional state obtain. So we will not be able to determine whether a mental state exists as it does because the o-physical properties are arranged in a physical state, or because they are arranged in a functional state.

It is plausible that metaphysical theories like Loar's and Stoljar's may rule out certain theories about which physical phenomena constitute which mental phenomena, but it seems clear that they do not themselves tell us what physical stuff constitutes what mental stuff. They do not close the epistemic gap.

3.3 Physical-Mental Identities and Conceptual Analysis

Block and Stalnaker argue more specifically against the concept of an epistemic gap. Their strategy is as follows. They argue that there is no a priori conceptual analysis which explains macrophysical truths in terms of microphysical truths. The example they use is of water and H2O. Even if we knew all the relevant physical facts about H2O, we still would not know that it is water. The mere conceptual knowledge of H2O and physical theory would not be enough to know that that stuff is the same as the clear, liquid stuff we call "water." But if we cannot know that H2O is water a priori, then we should not expect to know that certain brain states are certain mental states a priori. Of course, we do know that H2O is water. If we know that H2O is water without knowing that it is water a priori, then the fact that we cannot know that certain brain states are certain mental states a priori should not lead us to think that we cannot know those brain-mind identities at all. We should instead expect that the epistemic gap can be closed a posteriori, no unlike the gap between water and H20.

Here is another way to consider Block and Stalnaker's argument. Some philosophers argue that because we cannot know of physical-mental identities a priori, we cannot know of them at all. But if that is true, then it seems that we should be able to understand a priori those identities which we know to be true. We think that for any necessary identity, we should be able to come to know of that identity a priori, given that we are aware of the relevant facts. We know that there is a necessary identity between water and H2O. Block and Stalnaker claim that we cannot get from H2O to water a priori—that is, even if we are armed with the relevant facts about chemistry, physics, and so forth, we cannot simply reason out that the stuff we call "H2O" is necessarily identical

with the stuff we call "water." If we cannot know that H20 is water a priori, then we should not expect to know that certain physical states are necessarily identical with certain mental states a priori. In which case out inability to know of physical-mental states a priori is not a special feature of those identities. It should not lead us to think that there is a special epistemic gap in that case.

Block and Stalnaker's project covers a wide range of topics concerning conceptual analysis, but it will suffice for my purposes to focus only on their argument against the apriority of the H2O-water identity. David Chalmers and Frank Jackson wrote a response to this argument in which they give their reasons for believing that the H2O-water identity can be known a priori (and brain-mind identities cannot). I will explain their response and conclude along with Chalmers and Jackson that one can come to know the identity between H2O and water a priori, provided that one possesses the necessary facts. On the other hand, no background information will suffice to demonstrate that certain brain states are identical with certain mental states a priori. So there is a special epistemic gap in the case of physical-mental identities.

Block and Stalnaker's argument against the apriority of the H2O-water identity is as follows. Imagine a possible world in which the stuff we call "water" does not exist. On this "Twin Earth" the oceans are filled with a different substance—"XYZ"—which superficially resembles water and which is called "water" by the residents of Twin Earth. When XYZ is sufficiently heated, it does something that superficially resembles boiling but actually involves a different physical process. The residents of Twin Earth call this process "boiling," although it shares only superficial similarities with that process which we call "boiling."

Consider the idea of H2O and of boiling (as we use the term) from the perspective of the residents of Twin Earth. They would say "if H20 existed, it would not be water." And they would be correct, because "Water," as they use the term, does not refer to that stuff in our world which is H2O. Similarly, if the residents of Twin Earth were observing some XYZ boiling, they might say "if H2O were behaving like that, it would not be boiling." This would also be true, because they use the word "boiling" to refer to a different physical process than we do. If H2O were behaving in the way that XYZ behaves at high temperature, it would be doing something which only superficially resembles that process which residents of Twin Earth call "boiling."

Now let us further suppose that one resident of Twin Earth is provided with C, a complete microphysical account of a situation in which water is boiling, and T, a complete theory of physics. Block and Stalnaker tell us that this Twin Earthling could not deduce from T that H2O would boil under circumstances C. After all, they say, according to Twin Earthlings water cannot boil at all. H2O only does something which superficially resembles "boiling," as Twin Earthlings use that term.

Here Block and Stalnaker bring the argument back to our Earth. They ask us to imagine that a normal Earthling comes to know of C and T, but does not know anything else about how water boils (using our definition for both "water" and "boiling.") Could she deduce that water boils under circumstances C? No, say Block and Stalnaker, because she does not know that her situation is not like the one on Twin Earth. She does not know that the stuff in her environment is the same stuff which is the subject of the microphysical description C. As far as she knows, the stuff which she calls "boiling water" might be XYZ undergoing a physical process other than what we call "boiling." In

that case, the microphysical description C would not tell her anything about that the relevant stuff.

She might infer that the description C does refer to the relevant stuff in her environment, and that the description is a description of boiling water. In that case she would come to know that water boils under circumstances C. But that deduction would rely on additional information and therefore would not be a priori. But if she were to conclude that water boils under circumstances C on the basis of this additional information, she would have reached that conclusion a posteriori. She would not know that C refers to the relevant stuff in her environment a priori. Block and Stalnaker therefore conclude that there is no a priori conceptual analysis which leads us from the relevant microphysical facts (theory T and description C) to the relevant macrophysical facts (water boils under conditions C).

In their response to Block and Stalnaker's paper, Jackson and Chalmers argue that the above description leaves out certain information which would allow one to reach the relevant conclusion a priori. Specifically, they argue that the microphysical description C ought to be supplemented with locating information. Locating information is information that tells one what one's relationship is to the stuff in question. So if C were to be supplemented with locating information, it would consist not only of a microphysical account of a situation in which water is boiling, but also of information which describes one's own orientation toward the stuff in the account. As Jackson and Chalmers explain, the locating information adds a "you are here" sign to the microphysical account. A person armed with T and this modified version of C would know that her information describes stuff in her environment rather than stuff in some alien environment. She would

therefore be able to infer that C describes water, and that water would boil under circumstances C.

Jackson and Chalmers take this modified version of C to be the kind of description that matters in the relevant cases. When we analyze a particular physical phenomenon, we already know we know the orientation we have toward that phenomenon—it is the thing we are analyzing, the thing we have before us. So a conceptual analysis of that phenomenon will not only be based upon physical knowledge, but also implicit knowledge of our relationship to that thing. If we were to examine water in such a way that we could see individual molecules, we would not be in doubt as to the locational relationship of ourselves to the water or to the water molecules. It would be evident that the water and the water molecules have the same locational relationship to us—they are "that stuff" in front of us—and we could thereby conclude that the water and the water molecules are the same thing.

I will borrow a thought experiment devised by Jackson and Chalmers in order to drive this point home. Suppose we had a virtual reality helmet with provided detailed information about the world. A person wearing this helmet would have access to all the microphysical features of the phenomena around her. Moreover, the helmet would have the computing power to translate those microphysical features into specific data about specific objects, appearances, compositions, and so on and so forth. This helmet would therefore have all the microphysical facts, as well as the capacity to perform a priori calculations on the basis of those facts. It seems clear that the wearer of this helmet would know the macrophysical facts about objects around her. She would know, for example, that water is H2O. Therefore it is possible to arrive at microphysical to

macrophysical identities purely on the basis of the relevant physical facts and a priori reasoning.

On the other hand, it does not seem possible to arrive at physical-mental identities using this same method. Imagine that we were to put on the virtual reality helmet and take a look into another person's brain. The helmet would determine all the microphysical facts about the structure and chemical composition of the matter in the brain. And it would extrapolate from that microphysical data to tell us about the macrophysical structure and function of the brain's neural networks. Merely knowing about the brain's neural networks, however, would not tell us the phenomenal information about what that person is experiencing. We might know that a certain visual center of the brain is being stimulated, and that activity in that region of the brain is associated with a certain visual experience, but the helmet could not conclude *a priori* that there is any such visual experience occurring. The visual experience does not follow a priori from the physical information about neuronal structure and function.

But let us say that we *do* have such information. Let us suppose that we can flip a switch on the side of the helmet and turn on its mind reading function. Now the helmet not only tells us about the microphysical facts about the brain (and the macrophysical facts which follow a priori) but also about the phenomenal character of experience associated with that brain. We can see both the physical activity of the brain and the phenomenal experience that goes along with it.

I contend that even if we had access to this variety of information, the epistemic gap would remain. And the epistemic gap would prevent us from deciding between theories which differ with respect to how they describe the relationship of the physical to

the mental. Let us return to theories T1 and T2. According to T1, a mental state M will be identical with a given physical state P. And according to T2, a mental state M will be identical with functional state F. The information provided by the helmet will not rule out either T1 or T2, because whenever we observe M, the brain will also be in both a physical state and a functional state; so the physicalist can say that M exists as it does because of the brain's physical state P, and the functionalist can say that M exists because of functional state F. Human brains are always in a physical state and a functional state, so neither theory can be ruled out.

4 Looking for a Solution

There is an epistemic gap between the physical and the mental. We do not know how the mental arises from the physical, and we do not know what sort of physical phenomenon corresponds to a mental phenomenon. Of course, neuroscience can give us some answers. We know that the amygdala is important for the mental process of fear, that the hippocampus is important for memory, and so on and so forth. But we do not know the broader facts about the mental-physical relationship. Is activity in the amygdala associated with fear because a brain state in which the amygdala is active is identical with the mental state of being afraid? Or does amygdala activity play a certain role which is identical with that mental state? Or does activity in the amygdala give rise to a certain kind of behavior or behavioral disposition, and it is this behavior which is identical with fear?

If there is an epistemic gap, then no amount of physical information about the brain will help us determine which theory is true. Since there is an epistemic gap, we need some new source of information, or some new way of looking at the problem, in

order to determine which physical-mental identity theory is correct. In the proceeding chapter, I will introduce one such new way of looking at the problem.

Chapter 3: Self-Sampling and Consciousness

Human beings are conscious beings; we experience color, pain, taste, and so on. Apart from human beings, there may be other beings in our universe which are also conscious. It may well be that our universe contains some number of non-human beings that are behaviorally similar to humans: that is, they behave as though they are thinking, sensing, experiencing, and so forth. Of those beings in our universe, some proportion will actually have conscious experience.

In order to know how many of these non-human beings are conscious, we would have to know how many such beings exist, *and* what proportion of them is conscious. In this paper I am interested particularly in this second factor, for the following reason. Familiar views in the philosophy of mind, such as physicalism, functionalism, and

behaviorism, bear on the question of what proportion of beings are conscious. If, for example, everything that behaves as though it has conscious experience *actually does* have conscious experience, then all of the beings will be conscious. I will assume that there are other humanlike beings in the universe in order to simplify the proceeding discussion. But as I will explain later, my argument does not depend on this assumption.

I use a particular terminology for describing views which differ from one another with respect to what proportion of intelligent systems they take to be conscious. A **liberal** view is one which holds that a greater proportion of intelligent systems are conscious. A **conservative** view is one which holds that a lesser proportion of intelligent systems are conscious. I will argue that we have good reason to believe that a liberal view is correct. We therefore have good reason to think that a particular view in the philosophy of mind is correct insofar as it lends itself to greater liberalism.

My argument in favor of liberalism is as follows. If a liberal view is true, then it is more likely that human beings will be conscious. Human beings are conscious; therefore, we should assign a higher probability to a liberal view. In short, the fact that human beings are conscious is evidence that we can use to inform our judgments about the amount of consciousness in the universe as a whole. We can treat ourselves a sample of the total set of intelligent systems in the universe. For this reason, I refer to my argument as the **self-sampling argument**.

I will first demonstrate that the self-sampling argument is sound. I will then explore the ramifications of the self-sampling argument on traditional views in the philosophy of mind. In particular, the self-sampling argument is an argument in favor of certain metaphysical views insofar as those views lend themselves to liberalism. While

the self-sampling argument does not definitively prove that any particular metaphysical view is correct, it does provide *good reason* to believe particular views. It provides incremental support for some views over others.

1.1 Preliminaries

A few terms and concepts should be clarified. First of all, what do I mean by "being"? For the purposes of my argument, a "being" is an **intelligent system**, and an intelligent system is "intelligent" just by virtue of the way that it behaves. An intelligent system is anything which behaves as though it thinks, feels, experiences, and so on. My reasons for defining the word "being" in this way are essentially practical. If two views differ from one another with respect to what proportion of *non-intelligent systems* they take to be conscious, then the evidence of human consciousness will not help us decide between those views. Imagine one view holds that all physical systems are conscious, whereas another holds that all intelligent systems are conscious. Clearly, the first view holds that more stuff in the universe is conscious than does the second view. But the fact that human beings are conscious will not count as evidence in favor of one of these views over the other, because human consciousness is guaranteed under both views. Human beings are intelligent systems and physical systems. For the sake of simplicity, then, I describe those two views as equally liberal. A liberal view is one which is supported by the evidence of human consciousness. Liberality, then, must be defined in terms of what proportion of intelligent systems are conscious.

What do I mean by "conscious"? This term too requires clarification, because the word "consciousness" can correspond to a number of different concepts. In this paper, I use the word to refer to the experience one has as a conscious subject. If there it is

something it is like to be some kind of thing, then that thing is conscious. Human beings, for example, are conscious; there is something it is like to be a human being. We see, feel, experience, and so on. This concept of consciousness is sometimes called "phenomenal consciousness" or "P-consciousness."[7]

I take it to be at least logically possible that an intelligent system could fail to be conscious.[8] That is, is there is no *contradiction* in claiming that a given intelligent system is not conscious. Whether or not an intelligent system is conscious is an empirical question, because "intelligence" and "consciousness" are conceptually distinct. I defend this view elsewhere.[9]

One final point regarding intelligent systems and consciousness: we human beings are conscious intelligent systems, and we know it. Our own conscious experience is immediately accessible to us in a way that goes beyond our physical knowledge of our brain states, physical composition, and so on. We know we are conscious just by virtue of being conscious. Although some might be read as denying it[10] I do not believe this is a particularly strong or controversial stance on the nature of consciousness. I do not assert that mental states are essentially private or ontologically distinct from brain states. I am only asserting that our knowledge of our mental states goes beyond our knowledge of our own physical composition. It is generally accepted that this property of accessibility is attributable to our conscious states.[11]

2.1 Probability and Consciousness

---

[7] Ned Block (1998). *On a confusion about a function of consciousness.*
8 David Chalmers (1996). *Facing up to the Problem of Consciousness*
[9] <THESIS HERE>
[10] Daniel Dennett (1985). *Quining Qualia*
[11] David Rosenthal (2002). *Explaining Consciousness*

It is obvious that the fact of consciousness can be evidence for some hypotheses concerning the prevalence of consciousness in the universe. That we are conscious at least rules out the hypothesis that nothing in the universe is conscious. And it is plausible that one's consciousness is at least some evidence to believe that other humans are conscious.[12] What is potentially controversial is whether or not the fact that humans are conscious can be used as evidence for more specific hypotheses about the *prevalence* of consciousness in our universe. Is human consciousness evidence in favor those hypotheses that hold consciousness is relatively common in our universe? I will argue that it is.

For simplicity's sake, let us say that we are concerned with just two particular views in the philosophy of mind, Liberal, 'L', and conservative, 'C'. According to L, a greater proportion of intelligent systems in the world are conscious. According to C, a lesser proportion of intelligent systems in the world are conscious.

When we take into account the fact that human beings are conscious, we get some new evidence that bears on the disagreement between L and C. Human beings are an intelligent system, and human beings are conscious. There is a greater probability of humans being conscious if a greater proportion of intelligent systems are conscious, and a lower probability of humans being conscious if a lesser proportion of intelligent systems are conscious.

If $P(E|L) > P(E|C)$, then E is evidence in favor of L over C. Here I rely on a familiar Bayesian principle in epistemology, the **likelihood principle.** The likelihood principle tells us that new evidence counts in favor of a hypothesis if the evidence is more likely under that hypothesis. Put another way, if $P(E|H1) > P(E|H2)$, then E is evidence

[12] Bertrand Russell (1948) *The Argument from Analogy for Other Minds*

in favor of H1 over H2. According to the likelihood principle, E is evidence in favor of L over C.

The likelihood principle is a very plausible way to use new evidence to adjust one's conditional probabilities about competing hypotheses. Imagine the following scenario: suppose you are a birdwatcher and you are trying to determine whether a bird is a crow or a jackdaw. The two birds look similar, so at first you have no evidence for H1 (jackdaw) over H2 (crow). Eventually, however, you notice that the birds frequently share their food. You know that while jackdaws often share food with other jackdaws, crows only rarely share their food with other crows. So E, the frequent food-sharing behavior, is much more likely given H1, the hypothesis that you are observing a jackdaw rather than a crow. Therefore, you update your conditional probability about H1 and H2 by increasing the probability of H1 relative to H2. If the bird is sharing its food, than there is reason to believe it is a jackdaw rather than a crow. Thus, $P(H1|E) > P(H2|E)$.

When we use our own consciousness as evidence in favor of a liberal theory rather than a conservative theory, we are not doing anything terribly different than the birdwatcher. It is more likely for humans to be conscious if a greater proportion of intelligent systems are conscious, just as it is much more likely for the birds to be sharing their food if they are jackdaws.

Observing Human Beings 2.2

There are two general kinds of objections to the self-sampling argument. The first kind of objection has to do with the use of consciousness as evidence to support views about the prevalence of consciousness in the universe. The second kind of objection has

to do with using *our own* consciousness as evidence. I take this second kind of objection to be more serious.

I will respond to the first kind of objection by demonstrating that it can be admissible to use consciousness as evidence to support liberalism. I will do this by setting up a thought experiment in which entities other than human beings use human consciousness as evidence. This will suffice to show that consciousness can be evidence, without bringing in the issue of whether or not *we human beings* can use human consciousness as evidence. Once I have shown that it is admissible to use consciousness as evidence, I will move on to the more serious issue of whether or not is admissible to use *our own* consciousness as evidence.

Imagine that the evidence of human consciousness is being used by beings other than humans. Let us suppose that there are immaterial angels flying about the universe, trying to determine what is and what isn't conscious. These angels are immaterial, so the fact that they themselves are conscious does not help them determine what sort of physical beings might be conscious. The angels differ with respect to their beliefs about the prevalence of consciousness: one angel holds the liberal view 'L' while the other holds the the conservative view 'C.' So one angel believes that a greater proportion of intelligent systems are conscious, whereas the other believes that a lesser proportion of intelligent systems are conscious.

Suppose the angels access human consciousness just as easily as humans can access their own consciousness; when they do so, they find that human beings are in fact conscious. This new evidence bears on the angels' disagreement in a straightforward way: it is evidence in favor of the view which holds that a greater proportion of

intelligent systems are conscious. Humans were more likely to be conscious given that a liberal view is true than they were if a conservative view is true. If P(E|L) > P (E|C), and the angels discover evidence E, then E should increase their conditional probability in L over C.

This thought experiment seems to me to illustrate that one can use an entity's consciousness as evidence to support liberalism over conservatism in the way the self-sampling argument does. The angels use the evidence in an intuitively appropriate way: if one thinks that a lot of the intelligent systems in the universe are conscious, then of course one is bound to consider it more likely for a given intelligent system to be conscious. So finding out that some conscious system is intelligent confirms that theory to some extent.

It may seem strange in the abstract to use human consciousness as evidence, because it could be argued that *objectively speaking* the probability of humans being conscious is either 0% or 100%. If the laws of the universe are such that human brains give rise to or constitute consciousness, then the odds of humans being conscious is 100%. If not, then the probability is 0%. But I am not dealing in this sort of "objective probability."

I am dealing in subjective (or epistemic) probabilities. From our own subjective viewpoint, it makes sense to assign probabilities other than 0% or 100% even to outcomes that are determined objectively. Let us say that I have attempted to solve a complicated math problem. There is a sense in which the odds of my answer being correct are either 0% or 100%—either the numbers are right, or they aren't. But given that I am uncertain as to whether or not the numbers are right, it is perfectly sensible for

me to assign credence to my being correct that falls between 0% and 100%. Similarly, before we know the precise laws of the universe that would explain whether or not humans are conscious, it seems quite sensible to inform one's judgment based on the proportion of intelligent systems that are conscious. If we already *knew* the physical-mental relations that explain with certainty whether or not a given intelligent system is conscious, then of course we would have no reason to consider the proportion of conscious intelligent systems. But the angels do not know the physical-mental relations, and neither do we.

Using evidence of human consciousness does work for the angels. If it does not work for us, then it does not work because human beings cannot use human beings' own consciousness as evidence. In the following section I will examine a number of objections to the claim that we can use our own consciousness as evidence. These objections fail to show that it is impermissible to use this evidence.

***Observing Ourselves* 2.3**

The controversiality of using one's own consciousness as evidence stems from more general concerns about a traditional Bayesian model of conditionalization. One familiar problem with the Bayesian model is that it cannot incorporate known information as evidence for some hypothesis. If we are already know some evidence, then P(E) = 1 and P(H|E) = P(H). But if P(H|E) = P(H), then E does not raise the probability of H, so E is not evidence in favor of H. This is the **problem of old evidence**.[13]

If the problem of old evidence is a real problem, then it is a problem with the Bayesian model, not with the prospect of using known information as evidence. There are

[13] Clark Glymour (1980) *Theory and Evidence*

obvious cases in which our pre-existing evidence serves as evidence. Whenever scientists come up with a new theoretical model which explains existing phenomena, they are using those phenomena as evidence to support that theoretical model. The movement of stars through the sky is evidence of the heliocentric model of the solar system, even though we already knew about the movement of the stars before the conception of the heliocentric model.

There is a fairly simple way to dissolve the problem of old evidence. One simply considers the competing hypotheses as though one did not already know the evidence, and then adjusts one's conditional probabilities as if learning that evidence for the first time. This approach obeys the spirit of the Bayesian model and provides the correct result in the relevant cases.

Consider again the birdwatcher. Let us say that he has been watching the black birds for a few weeks now, and he knows they share their food frequently. That evidence is a known factor at this point, so P(E)=1. However, he has not learned that the food-sharing behavior is common among jackdaws and rare among crows. How should he adjust his conditional probabilities once he learns this new fact? It is obvious that the correct move is to assign a higher conditional probability to H1(jackdaw) and a lower conditional probability to H2(crow), just as in the previous iteration of the thought experiment. Nothing has changed except the order in which he received the relevant information. To achieve the correct result, then, the birdwatcher can put himself in the same epistemic position as he was in the previous iteration of the thought experiment: he can consider the evidence as though learning it for the first time. Before he knew the birds shared their food, $P(E) < 1$. The birdwatcher can "reset" his subjective probability

to P(E) < 1 by simply bracketing his knowledge of the food sharing behavior. When he does, he achieves the correct result: it is more likely that the birds are jackdaws than crows.

We can avoid the problem of old evidence in the same way as regards our own consciousness. We simply consider the way in which the evidence bears on the competing hypotheses as though we did not already know the evidence. We ask ourselves "what is the probability that human beings would be conscious, putting aside the fact that human beings are in fact conscious?" Clearly, if more intelligent systems in the universe are conscious, the probability that human beings would be conscious is higher. So the evidence bears on the competing hypotheses in the expected way: it supports liberalism over conservatism.

The problem of old evidence is not a serious objection to the self-sampling argument, but there are more serious objections. One possible objection is that our evidence is biased in a particular way. We are subject to the **anthropic bias**, which is the bias that we have on account of being extant, aware, thinking beings. If we were not extant, aware, thinking beings, then we could not know that we are conscious.[14] We have E, but we could not have ~E. Our evidence is therefore biased, and we cannot employ it in favor of a particular hypothesis.

To see the intuitive appeal of the objection from the anthropic bias, it is helpful to consider a classic example of selection bias. Let us suppose that a fisherman is wondering what size fish live in a nearby lake. To answer this question, he draws out a net full of fish from the lake, and sees that all the fish thus caught are big. Let 'B' be the hypothesis

that the lake contains only big fish, and 'M' be the hypothesis that the lake contains a mix of small and big fish. It seems that 'E' should be the evidence that the fisherman's net caught only big fish. To the fisherman, it seems straightforwardly obvious that P(E|B) > P(E|M), because he was more likely to only catch big fish if there only existed big fish to be caught. And so the fisherman concludes that B is correct: the lake contains only big fish. But the fisherman has made a mistake. Actually, his net contains holes too big to catch small fish, so the fisherman could not have discovered the evidence which would have supported M rather than B.[15] He could not have caught a net full of both small and big fish.

It seems natural to derive a principle about biases from this thought experiment. One plausible formulation of a principle might be:

Bias Principle: For two competing hypotheses H1 and H2, if you could not have had evidence to support H2, then your evidence cannot support H1.

This principle seems to explain the fisherman's mistake. His observation that the net is full of big fish does not support his conclusion that the lake is full of big fish, because there was never any chance of pulling up a net full of both big and small fish. In other words, he could not have had evidence to support M rather than B. Given that the fisherman could not have had evidence to support M, it seems intuitive that his observations cannot be evidence to support the competing hypothesis. If the fisherman could not have evidence to support M, he cannot have evidence to support B either.

If this bias principle is correct, then it arguably defeats the self-sampling argument. If we knew that we were not conscious, then we would have evidence in favor

[15] Eddington, *The Philosophy of Physical Science,* 1939

of a conservative view. But it is at least plausible that we could not have such evidence. One might argue that if we were not conscious, we could not have beliefs, and therefore could not learn anything whatsoever. I at least find this line of reasoning quite plausible.[16] But if we could not learn anything, then we could not learn that we were not conscious, and could not have evidence to support C over L. If we could not have evidence to support C, and if the bias principle is correct, then our observation that we are conscious is not evidence in favor of L.

However, it is not clear that the bias principle thus formulated is correct. Some other thought experiments present clear cases in which it is permissible to use the available evidence, even though one could not have had the opposing evidence. For example, Weisberg describes a case in which a prisoner is sent before a firing squad for execution. The prisoner knows that the firing squad will use one of two different kinds of guns: Type A guns, which are accurate, or type B guns, which are notoriously inaccurate. She knows that one of these two hypotheses will be correct: she will be shot with type A guns (H1), or she will be shot with type B guns (H2). The shots ring out, and the prisoner finds that she is unharmed. It seems natural for her to conclude on this basis that the firing squad used type B guns, because her survival was far more likely if they did.[17] But doing so is in violation of the bias principle as we have formulated it. The prisoner could not have known anything at all if she was killed by the firing squad, so she could not have had the evidence which would have supported H1. If we accept the bias principle, we must concede she ought not to count her survival as evidence to support H2. But this is obviously wrong. There must be some trouble with the bias principle.

[16] I argue this point in < THESIS HERE>
[17] Weisberg, *Firing Squads and Fine-Tuning,* 2005

Matthew Kotzen suggests a different way to check whether or not it is admissible to use certain evidence. His solution is as follows: when we compare the likelihood of some evidence under two hypotheses, we should check for both the likelihood of that evidence obtaining *and* for our likelihood of discovering that evidence. Call 'K' the means by which we gather our evidence. Kotzen suggests that in Eddington's fisherman example, the fisherman ought not to compare P(E|H1) to P(E|H2), but rather P(E&K|H1) to P(E&K|H2).[18] K, in the fisherman example, is the fisherman's means of catching fish with a large-holed net. So the fisherman should compare the likelihoods of catching a net full of big fish with a large-holed net under the hypotheses H1 and H2. If he does, he gets the correct result: it is equally likely to catch big fish with a large-holed net whether the lake contains only big fish or both big and small fish. Therefore he ought not to conclude that the lake contains only big fish.

Kotzen's method also returns the correct result in the firing squad thought experiment. In that scenario, K is the prisoner's continued existence (on the basis of which she has the evidence that she has not died), and E is the evidence that she has not died. Clearly, in this case, P(E&K|H1) > P(E&K|H2). The prisoner is more likely to survive, and more likely to know that she survived because she still exists, given that the executioners used the inaccurate guns. The bias principle is wrong. It can be permissible to use one's evidence in favor of some hypothesis, even when one could not have had evidence which supported the competing hypothesis.

Even though the bias principle is wrong, there is still clearly something to be learned from Eddington's fisherman thought experiment. The fisherman should not have concluded as he did that the lake was full of fish. There must be some lesson we can learn

[18] Kotzen, *Selection Biases in Likelihood Arguments,* 2012

from the thought experiment about when we cannot use the evidence available to us. To find this lesson, one need only look for the difference between Eddington's fisherman and Weisberg's prisoner. The fisherman is in the wrong, but the prisoner is clearly in the right. Neither the fisherman nor the prisoner could have had evidence to support a competing hypothesis—he could not have had a net full of small and big fish, and she could not have known that she'd been shot and killed. However, the two cases do differ in an important way. Whereas the fisherman could not have failed to gather the evidence he did, the prisoner *could* have failed to gather the evidence she did.

We have reached the crux of the issue. It is intuitive that if one had to have learned some evidence, one cannot use that evidence in favor of some hypothesis. But that is not to say that one must have been able to learn the competing evidence. The prisoner thought experiment illustrates that one can very sensibly use one's evidence even in cases where it would be impossible to have competing evidence. The original formulation of the bias principle was mistaken. It can be correctly reformulated thusly:

Revised Bias Principle: If you were equally likely to learn some evidence on competing hypotheses H1 and H2, then that evidence does not support one hypothesis over the other.

This principle captures the difference between the cases of the fisherman and the prisoner. The fisherman could not have failed to learn that he had a full net of big fish—he was equally likely to have that evidence whether the lake contained only big fish or both big and small fish. The prisoner, on the other hand, was not equally likely to know that she survived if she was shot with type A guns or type B guns. She was much more likely to be shot and killed if she was shot with type A guns, in which case she would not

know anything at all. This revised bias principle is the correct lesson we should draw from the Eddington's fisherman thought experiment.

We can incorporate the revised bias principle into the likelihood principle as follows:

Revised Likelihood Principle: Learning E supports H1 over H2 if you were more likely to learn E on H1 than on H2.

The revised likelihood principle takes into account the likelihood of the evidence obtaining, *as well as* the likelihood of our learning that evidence. It therefore incorporates our means of gathering evidence into the discussion, just as Kotzen suggests. Our means of gathering evidence bears on our likelihood of learning that evidence. Apart from incorporating that additional factor, the revised likelihood principle does not depart from the original. It works in just the same way, except that it requires that we take into account our means of learning the evidence in question. And it returns the correct result in every case. The fisherman was no more likely to learn E on B than on M, so E does not support B. The prisoner was much more likely to learn E on H2 than on H1, so E does support H2. And we were more likely to learn E on L than on C, so E does support L.

We are more likely to learn that we are conscious if a greater proportion of intelligent systems in the universe are conscious. If many intelligent systems in the universe are conscious, then a given intelligent system—namely, us—is more likely to be conscious. Therefore E is more likely on L than it is on C. It does not matter that we could not have the competing evidence. Of course we could not know it if we *weren't* conscious. What matters is that we are more likely to have the evidence on one hypothesis rather than another. And we are.

Self-Sampling and Modal Views 3.1

I now turn to the question of how the self-sampling argument bears on familiar views in the philosophy of mind. I am particularly interested in four materialist views: physicalism, functionalism, behaviorism, and panpsychism.

The materialist theses that interest me are often formulated as supervenience theses. The thesis of **functionalism** is that *mental states supervene on functional states*. Under functionalism, mental states exist as they do by virtue of the functional role they play, and can be realized by physically different systems. Functionalism maintains that intelligent systems that are alike with regard to functional organization are alike with regard to consciousness. If one kind of functional organization realizes or constitutes consciousness, then all intelligent systems with that functional organization will be conscious. In contrast, physicalism holds that the physical composition of a system matters, rather than merely its functional organization. Now sometimes 'physicalism' is used interchangeably with 'materialism', to refer to the weak view that systems that are alike physically are alike mentally as well. This sort of weak view is plausibly compatible with functionalism. But I use the term 'physicalism' to refer to a stronger view that holds moreover that functional similarity is not enough to guarantee mental similarity. So according to physicalism, *mental states supervene on physical composition*—and *not* upon functional states. Two systems can be functionally alike without being mentally alike, because one system may have the physical composition required for consciousness whereas the other does not. Therefore functionalism and physicalism are incompatible with one another. Another familiar view is **behaviorism**. Under behaviorism, *mental states supervene on behavioral dispositions*. Therefore any two systems that behave in the same way will also be mentally alike. All intelligent systems are by definition alike

with regard to their behavior, so according to behaviorism, all intelligent systems are alike with regard to consciousness. Those systems will be mentally alike even if they have vastly different physical compositions and functional organizations.

The final materialist view to be examined is somewhat less familiar than the others. **Panpsychism** is the view that *all matter whatsoever is conscious or proto-conscious in its essential character*, including even fundamental particles. Panpsychism maintains that however unintuitive it may seem, there is actually something it is like to be a fundamental particle.[19] Unlike the other materialist views, panpsychism is not typically formulated as a supervenience thesis. Because it is fundamentally different from the other views, I leave it out of the discussion until the end of the paper.

In order to show how the self-sampling argument bears on disagreements between these views, we must determine how liberal these views are relative to one another. That is, we must determine what proportion of intelligent systems each view takes to be conscious. If, according to one modal view, a greater proportion of intelligent systems are conscious, then that view will be more strongly supported by the evidence of human consciousness. One way in which one view can be more liberal than another view is by entailing that a greater proportion of intelligent systems are conscious. One view might entail that 100% of intelligent systems are conscious relative to some other view's 50%. On the other hand, views can be consistent with multiple hypotheses about the proportion of conscious intelligent systems, rather than entailing one specific hypothesis. One such view can lend itself to greater liberalism than another such view if it assigns higher probabilities to the hypotheses consistent with it. For example, two views might each entail one of two hypotheses: that either 0% or 100% of intelligent systems are conscious.

[19] Strawson, G. *'Realistic monism: why physicalism entails panpsychism'*

One of those views would lend itself to greater liberalism if it tended to favor the 100% hypothesis over the 0% hypothesis.

In this section I deal exclusively with materialist views; however, the discussion of these materialist views also bears on dualist views insofar as those views assign consciousness to the same proportion of intelligent systems. I take it to be possible that for every materialist view which holds that there is a necessary relationship between a given physical state and a given mental state, there is some dualist view that there is a mere nomological relationship between that physical state and that mental state. If a materialist view holds that a physical phenomenon *constitutes* a mental phenomenon, there is a corresponding dualist view which holds that the same physical phenomenon *causes or unerringly coincides with* the same mental phenomenon.[20] The upshot of those materialist and dualist views will be the same: the physical phenomenon goes hand in hand with the mental phenomenon. And for the purposes of the self-sampling argument, whether the relationship is metaphysical or nomological makes no difference—all that matters is that a given view holds that some proportion of intelligent systems are conscious. So although I deal explicitly with materialist views, the discussion of those views could be seamlessly substituted with a discussion of corresponding dualist views.

Modal Views and Liberalism 3.2

Now we must determine the relative liberality of the four views. It may seem obvious that the four views can be arranged in a certain way; namely, from behaviorism, to functionalism, to physicalism. This order of relative liberality may seem obvious for the following reasons. If human beings are conscious, and all intelligent systems are alike with regard to consciousness, then all intelligent systems are conscious. Therefore

[20] Chalmers, D. *Consciousness and its Place in Nature*

behaviorism is very liberal. It seems that there must be at least as many functional duplicates of human beings as there are physical duplicates, so more intelligent systems are conscious under functionalism as under physicalism. Therefore functionalism is more liberal than physicalism—though neither is as liberal as behaviorism. And it is hard to imagine that panpsychism could be conservative than any other view. At first blush, then, the order seems obvious: panpsychism to behaviorism to functionalism to physicalism.

This approach is problematic. The problem is as follows: we cannot use the fact of human conscious to demonstrate the liberality of a given view, because the fact that human beings are conscious is the evidence we use to support liberal views. We cannot "count" the evidence of human consciousness twice. Consider this: if we were not conscious (and somehow we knew that fact) then we would reverse the order given above. We would say that behaviorism is the most conservative, followed by functionalism, then by physicalism. This should suggest that there is some mistake being made when we use the fact of our consciousness to demonstrate that particular views are conservative or liberal.

Suppose a jar is filled with some combination of black and white marbles. There are two theories about the contents of the jar: T1 says that it contains either all black or all white marbles; T2 says that it might contain any combination of black and white marbles. We pick a marble out of the jar, and it turns out to be white. Now we might be tempted think something like the following. According to T1, the jar is full of white marbles. According to T2, it does not necessarily contain all white marbles. There are likely to be more white marbles according to T1 than there are According to T2, so picking a white marble is evidence in favor of T1 over T2. And here, we have already

picked a white marble! So T1 is more likely than T2. But clearly this is a mistake: we are counting the evidence of the white marble twice. We must avoid making the same mistake about the fact of our consciousness. In the proceeding discussion it will become more clear where the risk arises, and how it can be avoided.

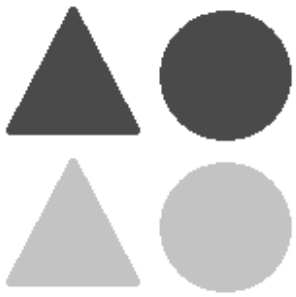
Despite the fact that we must bracket our knowledge of human consciousness, I believe it is possible to demonstrate that some modal views are more liberal than others. In the following sections, I will explore this possibility. I will first compare physicalism and behaviorism, then functionalism and physicalism. Although each modal view permits of being formulated in multiple ways, it is possible to draw some rough conclusions about the relative liberality of the modal views to each other.

***Behaviorism and Physicalism* 3.3**

To determine the liberality of a modal view, we must see how it lends itself to liberal and conservative hypotheses. We must identify the hypotheses which are consistent with each modal view to see if the set of hypotheses entailed by one view is more liberal than those entailed by another. The view which lends itself to greater liberalism will be the one which tends to assign higher probabilities to more liberal hypotheses.

One way to identify the hypotheses consistent with a given modal view is to bring that modal view to bear on an actual set of intelligent systems. Of course, we do not know the actual set of intelligent systems in our universe. But we can invent a set that will work for our purposes. Let us imagine that the universe contains a total of only four intelligent systems. Each of these systems differs from the others with respect to its physical composition. I will use a set of four symbols to represent those intelligent

systems: a light triangle, a dark triangle, a light circle and a dark circle (the shapes and colors will matter later on). The symbolic representations of the four intelligent systems are below:



Let us bring the modal view of physicalism to bear on this set of intelligent systems. Given that physicalism is true, there are sixteen possible hypotheses about which of the four intelligent systems are conscious. It is consistent with physicalism for any possible combination of intelligent systems to be conscious. One hypothesis is that only the light circle is conscious, another is that the dark triangle and the light triangle are conscious, another is that all of except dark circle are conscious, and so on. Assuming for now that we have no reason to favor one of these hypotheses over another, on average these sixteen hypotheses will hold that two of the four intelligent systems are conscious.

Given that behaviorism is true, there are only two possible hypotheses about which of the systems are conscious: either all of them are conscious, or none of them are. The systems are all alike with regard to their behavior—they all behave intelligently—so they must be alike with regard to consciousness. The extreme liberal and the extreme conservative hypotheses average out in the same way as do the physicalist hypotheses—on average, two of the four systems are conscious.

Of course, the fact that human beings are conscious allows us to rule out the extreme conservative hypotheses according to which no intelligent systems are

conscious. However, for reasons that have already been discussed, we cannot use that evidence at this juncture. The evidence of human consciousness has to be "saved" for the self-sampling argument. In determining the relative liberality of modal views, we must bracket our knowledge of human consciousness; we will then bring that knowledge to bear in demonstrating that the more liberal views are more likely to be correct. Bracketing our knowledge of human consciousness, the physicalist and behaviorist sets of hypotheses average out the same way. Both of the sets hold that on average, two of the four intelligent systems are conscious.

It may be tempting to conclude on that basis that physicalism and behaviorism are equally liberal. If we bracket our knowledge of human consciousness, and we distribute our conditional probabilities equally among all the hypotheses available to us under both physicalism and behaviorism, then it does appear that they are equally liberal. Under behaviorism, we would say that there is a 50% chance that all intelligent systems are conscious, and a 50% chance that none of them are conscious. Under physicalism, we would say that there is a 6.25% chance that any one of the sixteen possible physicalist hypotheses is true. In either case, we would expect that the odds of a particular intelligent system being conscious are 50%.

However, it is unclear that we should distribute our conditional probabilities evenly among the available hypotheses. One reason to think that we should not distribute them evenly is that our reasons for believing the extreme conservative physicalist view are better than our reasons for believing the extreme conservative behaviorist view. Consider the reasons we may have for believing the extreme conservative hypothesis on the assumption that physicalism is true. We might think that some intelligent system

could be conscious, if it existed, but that none of the extant conscious systems are conscious. Perhaps we think that only black squares are conscious, for example. On the other hand, we might think that no physical thing whatsoever could be conscious. This hypothesis is also consistent with physicalism. If nothing physical could be conscious, then everything that is alike with regard to its physical composition also must be alike with regard to whether or not it is conscious.

Under behaviorism, on the other hand, there is only one reason for believing that no intelligent systems are conscious. As with physicalism, we can hypothesize that no physical thing, or no intelligent system, could be conscious. However, we cannot hypothesize that although some intelligent system *could* be conscious, none of the existent intelligent systems are. The intelligent systems are by definition alike with regard to their behavior: they all behave intelligently. We cannot hold that some intelligent systems are conscious and others are not, because under behaviorism we must hold that all intelligent systems must be alike with regard to whether or not they are conscious. Of the two reasons we can have for believing the extreme conservative hypothesis under physicalism, we can only have one of those reasons under behaviorism.

Say B_0 is the hypothesis that behaviorism is true and no intelligent systems are conscious, and P_0 is the hypothesis that physicalism is true and no intelligent systems are conscious. It seems that we must assign at least as high a conditional probability to P0 as we do to B0, since we have the same reasons for believing P_0 as we do for believing B_0, plus further additional reasons. If we simply distributed our probabilities equally across all the hypotheses available to us under physicalism and behaviorism, we would do the opposite. We would assign a much higher probability to B_0 than to P_0. But that must be

wrong, because the physicalist has the same reasons for believing the extreme conservative hypothesis as does the behaviorist, *plus* additional reasons.

Let us say that we must assign at least as high a probability to P_0 as we do to B_0. When we assign equal probabilities to P_0 and B_0, respectively, and distribute the remainder of the probability equally to the remaining hypotheses, behaviorism turns out to be significantly more liberal than physicalism; it seems we should believe that a given intelligent system is twice as likely to be conscious under behaviorism as it is under physicalism. That is, the modal view of behaviorism lends itself more toward liberal hypotheses than does physicalism. The exact conditional probabilities will depend upon how much probability we assign to the hypotheses that nothing could be conscious. But as long as we assign an equal probability to the extreme conservative hypothesis under physicalism and behaviorism, behaviorism turns out to be more liberal than physicalism.

We can return to the example of the philosopher angels to illustrate why the behaviorist assigns greater probability to the proposition that a particular intelligent system is conscious. Suppose that some angels are flying about the universe and happen upon human beings. They see that these human beings are intelligent systems. The angels try to decide whether or not they are in fact conscious, and this sparks a debate between them. The angels agree in that they are in doubt as to whether or not any physical thing in the universe could be conscious at all. But they disagree about what in the universe would be conscious, if anything could be conscious at all. One claims that if anything physical were to be conscious, it would depend on whether or not it behaved intelligently. The other claims that if anything physical were to be conscious, it would depend upon the specific physical composition of the intelligent system in question.

When the angels peer into the minds of human beings and discover that they are in fact conscious, this new information will count as evidence in favor of the behaviorist angel. The angels were equally in doubt as to whether or not any intelligent systems at all could be conscious. Putting that doubt aside, the behaviorist angel was sure that any given intelligent system would be conscious. The physicalist angel was not sure. Intuitively, then, the behaviorist angel thought it more likely that human beings would turn out to be conscious. The behaviorist angel had only one reason for thinking that humans might not be conscious; the physicalist angel had two. When the angels discovered that human beings were conscious, they gained new reason to believe that behaviorism is correct.

Behaviorism is therefore more liberal than physicalism, even bracketing knowledge of our own consciousness. Now I will turn to functionalism.

Functionalism 3.4

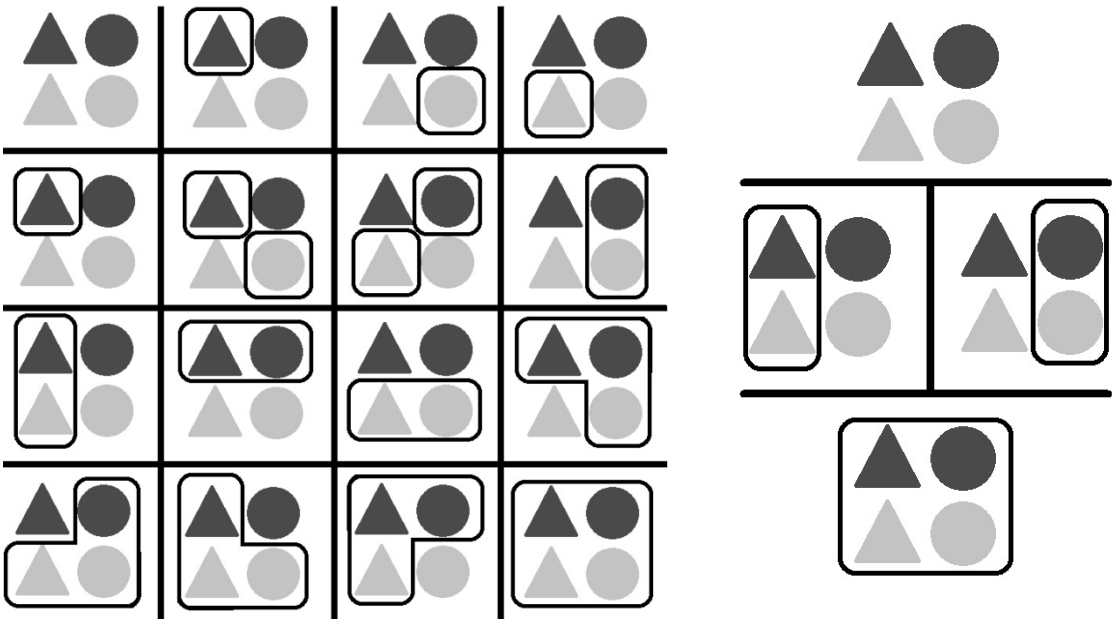
Intuitively it seems as though functionalism should sit somewhere between physicalism and behaviorism. One gets the sense that functionalism sets looser conditions for consciousness than does physicalism, but not as loose as those set by behaviorism. Functionalism does not require that an intelligent system have a particular physical composition, but only that it has a particular functional organization. However, unlike behaviorism, functionalism does not hold that *all* intelligent systems are conscious if any of them are. Like physicalism, it is consistent with the possibility that some intelligent systems are conscious while others are not.

I will argue that this intuitive sense is misleading. Actually, if we follow the same reasoning as we did with regard to physicalism and behaviorism, it seems that

physicalism and functionalism are equally liberal. That is not to say that functionalism *is* as liberal as physicalism. It is only to say that the argument which demonstrated that behaviorism is more liberal than physicalism will not demonstrate that functionalism is more liberal than physicalism. It may turn out there exists some other argument which demonstrates that functionalism is more liberal than physicalism; I address this possibility later on.

To demonstrate why functionalism seems to be as liberal as physicalism, let us return to the four symbols. Let us say that the symbols of the same shape represent intelligent systems with the same functional organization. According to functionalism, intelligent systems that are the same with regard to functional organization will also be the same with regard to whether or not they are conscious. Assuming that functionalism is true, we cannot hold both that the dark triangle is conscious and that the light triangle is not, for example. The set of functionalist hypothesis will therefore be more restrained than the set of physicalist hypothesis. Those sets of hypotheses are illustrated below.

Physicalist Hypotheses **Functionalist Hypotheses**



I have argued that the set of hypotheses under physicalism will contain one hypothesis according to which nothing physical could be conscious in principle. This hypothesis is consistent with functionalism as well. We can be functionalists and still speculate that no intelligent system whatsoever could be conscious. This hypothesis is consistent with functionalism as I have defined it (and how it is normally defined). Moreover, the functionalist will have the same reasons for believing that the extreme conservative hypothesis might be true as does the physicalist. That is, they can maintain that no intelligent system *could* be conscious, or they could maintain that although an intelligent system could be conscious, no existing intelligent system actually is conscious.

If we do not have better reason to believe F_0 than we have to believe P_0, then we cannot establish that the set of functionalist hypotheses is more liberal than the set of physicalist hypotheses in the same way that we established that the set of behaviorist hypotheses is more liberal than the set of physicalist hypotheses. We have no obvious reason not to distribute probabilities equally among all the available hypotheses, and the averages of the remaining hypotheses in each set are equal to one another. There are four functionalist hypotheses to physicalism's sixteen hypotheses, but the math works out the same. In this example set of intelligent systems, the conditional probability one should assign to the proposition that a given intelligent system is conscious should be the same whether one assumes physicalism or functionalism. Therefore it seems that physicalism and functionalism are equally liberal.

It may seem a bit odd that functionalism should not be more liberal than physicalism. We have already stated why functionalism seems intuitively to be more liberal than physicalism: it seems to set looser requirements for consciousness than does

physicalism. We might also have another reason for thinking functionalism is more liberal. Besides the intuition that functionalism sets looser requirements, a cursory examination of the relationship between physically and functionally identical intelligent systems seems to demonstrate that functionalism *must* turn out to be more liberal than physicalism. Every physical duplicate of an intelligent system is also a functionalist duplicate, but not vice versa. So if we know that one intelligent system is conscious, we must also know that at least as many other intelligent systems will be conscious under functionalism as will be intelligent under physicalism. According to functionalism, all of the intelligent system's functional duplicates will be conscious. According to physicalism, all of the intelligent system's physical duplicates will be conscious. Since the set of functional duplicates is at least as large as the set of physical duplicates, it seems as though functionalism is at least as liberal as physicalism, and possibly more so.

This argument is unsound because it does not account for the fact that under functionalism, we are just as often committed to maintaining that some intelligent systems *are not* conscious. Functionalism is just the view that things which are alike with respect to their functional states will also be alike with regard to their mental states. Functionalism therefore does not only tell us that functional duplicates of conscious systems will also be conscious; it also tells us that functional duplicates of systems that are *not* conscious will also not be conscious. Whenever the physicalist claims that an intelligent system is conscious, but its functional duplicate is not, the functionalist must disagree. But the functionalist need not disagree by claiming that *both* systems are conscious. It is entirely in keeping with functionalism to claim instead that *neither* system is conscious. So it is not clear that whenever a physicalist hypothesis is incompatible with

functionalism, the functionalist version of that hypothesis will include more conscious systems. If the functionalist hypothesis cannot be the same as the physicalist hypothesis, then there will be more than one corresponding functionalist version of that hypothesis. The functionalist can alter the set of conscious intelligent systems by claiming either that more intelligent systems are conscious, or that fewer are.

It is worth noting that actual functionalists might not *want* the support of the self-sampling argument. Functionalists often write as though the functional organization of the human brain is *a priori* constitutive of consciousness.[21] If the mere coming together of information from various inputs in the brain is *a priori* constitutive of consciousness, then the functionalist probably does not think she needs the kind of evidence that the self-sampling argument provides. That sort of functionalist thinks that functionalism is already provably true based on existing evidence. But functionalists who think that functionalism is true *a posteriori* have reason to look for more evidence in favor of their view. Such functionalists should be interested in *a posteriori* evidence of the sort that the self-sampling argument provides.

It is also worth noting that the liberality of functionalism may depend on how broadly the functionalist is willing to attribute "functional similarity". If the term "functional duplicate" is construed very loosely, it may turn out that all intelligent systems are functional duplicates of one another, in which case there will only be two possible hypotheses will be possible: either all intelligent systems are conscious, or none of them are. For example, according to the integrated information view, all systems are conscious insofar as they integrate information in some way.[22] It seems exceedingly

[21] David Chalmers discusses such views in his *Consciousness and its Place in Nature*
[22] Tononi, *Consciousness as integrated information: a provisional manifesto.*

likely that all intelligent systems must integrate information to some extent; otherwise they would not be "intelligent." In that case, all intelligent systems will necessarily be alike with regard to consciousness, and the integrated information view will be as liberal as behaviorism, and for the same reasons. We should divide our conditional probabilities between only two hypotheses, just as we did under behaviorism, and therefore such a formulation of functionalism is as liberal as behaviorism.

There is another reason why functionalism might turn out to be more liberal. We might also find out that we should not distribute our conditional probabilities evenly among all the possible functional hypotheses. Perhaps we should believe that if functionalism is true, some of the hypotheses entailed by functionalism are more likely than others, even setting aside that humans are conscious. A given functionalist view may maintain that although we do not know *a priori* that certain physical systems are necessarily conscious, we still have good *a priori* reasons to think that they are. Or at least, we have *a priori* reasons to think that some functional systems are more likely to be conscious than are other functional systems. The extent to which a formulation of functionalism is liberal will therefore depend in large part upon the *a priori* considerations which shape how the view should assign conditional probabilities.

Panpsychism 3.5

Panpsychism, unlike the physicalism, functionalism, and behaviorism, is not usually formulated as a supervenience claim. Panpsychism is simply the view that all matter whatsoever is conscious. Unlike behaviorism, functionalism, and physicalism, there is no extreme conservative hypothesis of panpsychism—but only because no one takes seriously the hypothetical conservative opposite of panpsychism. No one takes

seriously the "nullpsychist" view, according to which no matter whatsoever could be conscious. The reason that this view is not taken seriously is because we human beings are conscious. But we are bracketing our knowledge of human consciousness, so we should take it seriously. Just as we considered the extreme liberal and extreme conservative versions of behaviorism, so too should we consider both panpsychism and nullpsychism.

What might it mean for all physical matter to be conscious? It quite plausibly means that all intelligent systems are conscious (along with everything else, of course.) Let us assume for our purposes that the panpsychism entails the following thesis: *all intelligent systems are alike with regard to consciousness*. If that thesis is true, then panpsychism is as liberal as is behaviorism. Moreover, panpsychism is better positioned to be supported by evidence from the self-sampling argument. Behaviorists often think that all intelligent systems must be conscious because they believe behavioral dispositions are *a priori* constitutive of mental states, just as functionalists sometimes hold that the functional organization of the human brain is *a priori* constitutive of consciousness. It is harder to believe that matter itself is *a priori* constitutive of consciousness. And it is hard to imagine how any scientific investigation could determine whether or not fundamental particles have conscious experience. Panpsychism is therefore more likely to embrace incremental support which the self-sampling argument provides.

Panpsychism is a difficult theory to swallow, because it is very hard to take seriously the idea that protons and electrons have any kind of conscious experience whatsoever. But perhaps we will have to learn to take it seriously. Suppose that we have

good reasons to reject other strongly-liberal theories, such as behaviorism and integrated-information functionalism. Then we may find ourselves in the startling position of having to choose between two highly unattractive hypotheses: either panpsychism is true, or the fact of human consciousness is a coincidence on a cosmic scale. As unintuitive as it may seem, panpsychism may turn out to be the least strange hypothesis available to us.

Closing Comments 4.1

Human beings are more likely to be conscious if consciousness is more prevalent in the universe. Views in the philosophy of mind bear on the prevalence of consciousness in the universe. Therefore the fact that human beings are conscious bears on the question of which view in the philosophy of mind is true. I have identified a few views that are supported by the evidence, such as behaviorism, integrated-information functionalism, and panpsychism. However, I have not attempted to determine the prior probabilities of these views. If we knew both which views are supported by the evidence, and which views were most independently plausible, we could determine which views are on the whole most likely to be true. Determining the prior probabilities of these views is therefore a promising project.

Other future projects can be aimed at the self-sampling argument itself. For the purposes of this paper, I have framed liberalism and conservatism in terms of the proportion of intelligent systems they take to be conscious. But perhaps there are no other intelligent systems in the universe. It seems like the number of intelligent systems in the universe should not matter for the self-sampling argument. What matters is how the *prevalence* of consciousness among (however many) intelligent systems bears on our epistemic attitudes. But if we imagine that we are the only intelligent systems

whatsoever, then it is not obvious how liberalism and conservatism ought to be defined. We could alter the definition of liberalism such that it is a view about the prevalence of consciousness among a *hypothetical* pool of intelligent systems. But then, how should we populate this hypothetical pool? Should we take into consideration that some kinds of intelligent systems are more likely to exist, and thus should be better represented in the pool? Alternatively, we could define liberalism as a view about the likelihood of a *particular* intelligent system being conscious, but this approach suffers from similar problems. Let us say that human beings could have evolved other than they did, and that liberal views are views which hold that we were more likely to evolve to be conscious. How should we populate the hypothetical pool of possible ways in which humans could have evolved? Should some more likely evolutionary tracks be better represented?

In any case, I believe the foregoing argument gives us a good picture of how our own consciousness bears on familiar debates in the philosophy of mind. Discussions of consciousness are often preoccupied with the question of *how it could arise in us.* But there are multiple explanations to this question, each of which is *prima facie* compatible with our evidence about the physical world and our consciousness. We need some way of determining which of these explanations is most likely to be true. The self-sampling argument demonstrates one such method.