Mentality and Modality

John P. Burgess
Department of Philosophy
Princeton University

200609
for Ruby and Stella
“Material things I think, or feel, or see;
   All else is immaterial to me.”
— Ambrose Bierce, Devil’s Dictionary
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preface</td>
<td>iv</td>
</tr>
<tr>
<td>1 Descartes and Dualism</td>
<td>1</td>
</tr>
<tr>
<td>2 Functionalism and Supervenience</td>
<td>25</td>
</tr>
<tr>
<td>3 Epiphenomenalism and Causation</td>
<td>46</td>
</tr>
<tr>
<td>4 Physicality and Phenomenality</td>
<td>72</td>
</tr>
<tr>
<td>5 Possibility and Reference</td>
<td>94</td>
</tr>
<tr>
<td>6 Reflection and Projection</td>
<td>124</td>
</tr>
</tbody>
</table>

References | 152 |

Index     | 157 |
Preface

This short book is on a topic familiar to all students of philosophy, and a subtopic familiar at least to specialists in the philosophy of mind, but it is written from an uncommon point of view. The topic is the mind-body problem, now almost four centuries old, going back as it does to René Descartes and Cartesian dualism. The subtopic is the thesis of the supervenience of the mental on the physical, under discussion for about half a century now since it was first adumbrated by Donald Davidson to be later elaborated by Jaegwon Kim. The point of view from which these topics are approached is that of an outsider come to the issues from modal logic rather than philosophical psychology.

The main conclusions to be suggested or insinuated in this work, which might have been subtitled “Suspicions about Supervenience,” are two: first, that the supervenience thesis is false; but second, that whether it is true or false is nowhere near as important as has often been assumed. If these conclusions were accepted, one side in a long- and hard-fought contest would first be awarded the victory trophy, then told it is not gold but gilt. It would be unrealistic to hope to settle issues so long in dispute as those addressed here in so slim a volume as this, but I hope with the contrasting pair of conclusions just enunciated I may provide an irritant stimulus and do something to shake up the current state of debate.

In the text I have generally written as if addressing a newcomer, not previously acquainted with the questions under discussion. Why such a faux naïf approach? It comes down to this, that in looking at the literature of the subject from my outsider’s perspective, I found myself on quite a few points inclined to doubt what seemed to be widely assumed among insiders. I was shocked, for instance, at how many seem confident that it makes sense to attribute propositional attitudes to zombies. I thus implicitly have to ask for a good deal in the way of suspension of disbelief from readers who might come to this book after having
steeped themselves in the mainstream literature and absorbed its background assumptions. Perhaps the only practicable route to the required suspension is through projecting oneself imaginatively into the situation of one encountering the issues for the first time. It is to assist such an effort of imagination that in the text I do not from the beginning assume, as I have in this preface, that the reader has already heard of dualism, supervenience, zombies, and so forth. In the text, any usage in principle not assumed familiar will appear in **boldface** where it is first explained in what sense I am using it.

This work is in part a sequel to my *Saul Kripke: Puzzles and Mysteries* (and the concluding chapter incorporates some related material from my Kripke Lecture at the CUNY Graduate Center in 2012), though no familiarity with my earlier work is presupposed. My general intellectual debts remain those of the earlier book, now joined by the influence of key works of Ned Block, David Chalmers, Frank Jackson, Thomas Nagel, and (more in privately-circulated than published items) Stevan Harnad. I am grateful to Oliver Marshall, Neil Tennant, Alexander Williams, and Alexi Burgess for opportunities to present portions of this material publicly as it developed. I am grateful to the hosts and audiences at those public presentations, and the five leaders in the field just listed, and my colleagues Daniel Garber and Mark Johnston and former student Jack Woods, not to mention anonymous referees, for comments at greater or lesser length on earlier versions of parts of this work. It was a cryptic remark of my late colleague David Lewis, to the effect that virtually *all* present-day positions in philosophy of mind would have been considered by Victorians to be forms of materialism, that first inspired me to undertake this study.

Princeton
Spring, 2020
1 Descartes and Dualism
1.1 Substance Dualism

Fifty years ago a number-theorist writing about the problem of representing an $n$th power as a sum of two $n$th powers and a philosopher of mind writing about the problem of the relation of the mental to the physical might well have expressed themselves in similar terms. “Work on the problem,” such a commentator might have said, “constitutes only a restricted and rather specialized part of a large and growing field. But given the venerable age of the question, first brought to the attention of the learned world by a seventeenth-century French thinker,” and here the allusion would be to Pierre de Fermat in the mathematical case, René Descartes in the philosophical case, “it will always have a certain romantic appeal, drawing to itself some of the most acute thinkers in each generation — and inevitably also any number of cranks.” A writer addressing the philosophical problem might do so in similar terms today; by contrast, the mathematical problem is now solved, and the prize long on offer for its solution collected. Philosophical problems are more refractory than mathematical problems, some might conclude.

Virtually any expository writer addressing the mind-body problem would mention Descartes, as I have just done. Many would make a point of mentioning also what have been perceived as similar or related problems in other intellectual and cultural traditions. But it is to Descartes and a handful of his contemporaries that we primarily owe the two features most distinctive of our mind-body problem: first, the neural orientation, emphasizing the nervous system, especially the brain, and not the heart or liver or spleen or kidneys, as the part of the body most relevant to mental life; second, the mathematicization of the concept of the material or physical that is contrasted with the mental, arising from the seventeenth-century refocusing of science on what is quantifiable to the exclusion of what is not, for instance, on measurable size and shape, to the exclusion of sensed color. These two
features are so much taken for granted today that there is a danger of forgetting that they once were contentious novelties.

A striking feature of recent discussion of mind and body is that much of it is concerned with zombies. To be sure, no large number of present-day philosophers are believers in such beings: perhaps a century or so ago, in the days of William James and the Society for Psychical Research, there may have been philosophers or psychologists who seriously endorsed claims of their existence, but this is hardly so today. What is debated today is not their actual but rather their possible existence. The main topic of this book will be a third feature that the contemporary mind-body problem shares with Descartes’ of four hundred years ago, namely, the curious entanglement of the issue mentality and physicality with the notion of possibility — or with modality, to give the category containing such notions as possibility and necessity its usual name. I will in the end plead for disentanglement, but to make a case for disentanglement is only my final destination, and I attach less importance to arriving at it than to journeying towards it, enlarging our understanding of mentality and modality along the way.

While a writer on the mind-body problem in a handbook or encyclopedia will doubtless mention Descartes, if working under any halfway serious word limit such a writer will be unlikely to linger long over the historical background. Typically, the exposition will fast-forward to the present day, with the remark that the question is no longer debated in the same terms in which it was addressed four hundred years ago. In jargon, it is usually said that while Descartes advocated a “substance dualism,” even his closest followers today advocate only a “property dualism.” I will need to stay just a bit longer with Descartes here, though owing to limitations of space and expertise I will have to leave aside all scholarly exegetical debates, and present a portrait sometimes simplified to the point almost of caricature.
Our direct interest will be in human mental life, that of living, embodied members of our species, though of course the existence or non-existence of a number of the other kinds of mental lives has been debated from long before Descartes’ day. To begin with, there is the question of ghostly mental life, that of deceased, disincarnate ex-members of our species, and then several other cases: animal, robotic, extraterrestrial, angelic. The existence of ghostly minds, which is to say, the occurrence of survival after death, is the case most central to Descartes’ most celebrated philosophical work, *Meditations on First Philosophy*, and the only case that will be seriously considered here. (No one today will want to defend Descartes’ views on animals, even if no credence is given to the tale that he once kicked a dog and dismissed the resultant yelping as the mere squeaking of gears in an unfeeling machine.)

Already in Descartes’ day a domain of philosophy, which unlike theology may make no appeal to faith, revelation, tradition, authority, or the like, but only to reasoned argument and the evidence of experience, had for centuries been recognized in principle. But in practice arguments that passed for philosophical were often directed towards proving conclusions dictated in advance by theology; and Descartes’ work on mind and body, as he himself presents it, is a case in point. For he professed a faith according to which the mind or soul survives death and exists in a disembodied state pending resurrection and re-embodiment; and he maintained that this doctrine is not merely something true and known by faith, but something that could and should be proved philosophically as part of a defense of that faith. Yet while he claimed survival after death to be philosophically provable, in his *Meditations* he contents himself with the more modest goal of arguing for the weaker conclusion that, whether or not disembodied minds actually do exist, they possibly could. In Descartes it is with this claim that modality comes in.

Moreover, it is in the argument for this claim that talk of substance comes in,
as part of a quaint apparatus of substance *versus* mode and essence *versus* accident that sometimes makes the seventeenth century seem closer to the thirteenth than to the twenty-first. For Descartes’ argument, in executive summary so to speak, consists of claiming a clear and distinct perception of the truth of three premises which together yield the possibility of mind without body as an immediate consequence: first, that matter is a substance whose essence is extension, not thinking; second, that mind is a substance whose essence is thinking, not extension; and third, that when substances are distinct in essence it is possible for one to exist without the other, which is to say, for God to create one without the other, or having created both, to preserve one without the other.

Since the term “mode” has vanished from the vocabulary of philosophy over the last several centuries, and “substance” as Descartes uses the term only makes sense in contrast with “mode,” it is not incorrect to say that few present-day philosophers believe in mental substance in a Cartesian sense. But it is misleading, since few believe in material substance in a Cartesian sense, either. The usage of “substance” in present-day philosophy simply is not Descartes’ quasi-scholastic usage, and hardly differs from usage outside philosophy, where a *substance* is a kind of stuff as opposed to a kind of thing, the denotation of a mass noun like “gold” or “water,” rather than a count noun like “wedding ring” or “ice cube.”

Contemporary philosophers do not believe in mental substance in such a sense, but then neither did Descartes. For it is a characteristic of substances in this sense to be extended, to take up space, while this is just what minds do *not* do, according to Descartes. There have been believers in space-filling mental stuff, but Descartes was not of their number. They have included, rather, traditional thinkers who took spirit to be literally a kind of breath, and some in the days of James who were duped by mediums claiming to produce what they called “ectoplasm,” the paradigmatic mental or spiritual substance.
What is really meant by the claim that substance dualism is not the issue today is something that would be less misleadingly expressed avoiding the term “substance” altogether. It is that contemporary dualists are less concerned to argue that a human being consists of two irreducibly different components, a human body and a human mind, than that human life has two irreducibly different kinds of aspects, physical and mental. Talk of “the mind” may be understood as amounting to little or nothing more than a way of talking collectively about the ensemble of mental aspects.

And what is meant by saying that property dualism is now the focus of debate, as we are often told? This formulation suggests that there is a consensus that what I just now nebulously called mental and physical “aspects” should be understood as mental and physical properties. But this suggestion is false: there is no such consensus. Some influential figures have been deeply suspicious of the very notion of properties and have emphatically maintained that the important issue is that of mental and physical events. Elsewhere in the literature the key term may be “fact.” My own preference would be “states,” or where change is involved, “processes,” and I will mainly stick to those terms, though sometimes it will be more appropriate to adopt the chameleon-like policy of using at a given point the most common terminology used in the literature on the issues under discussion at that point.

Here note must be taken of a distinction between general types or patterns and particular tokens, their instances, giving us eight categories: being angry is a mental state type; having excess bile is a physical state type; becoming angry is a mental process type; acquiring excess bile is a physical state type; my being angry yesterday noon is a mental state token; my having excess bile yesterday noon is a physical state token; my getting angry yesterday noon is a mental process token; my acquiring excess bile yesterday noon is a physical process token. The term
“event” generally connotes a token, “property” a type, though one may speak also of event-types and property-tokens, the latter often called “tropes.” But “state” may seem more ambiguous, and below, if and where it makes a difference, types and not tokens should be understood unless otherwise indicated.

Before taking leave of the old-fashioned substance-mode distinction, I should mention another way it figures in Descartes. His most fundamental insight is that, delusion being a form of (erroneous) thinking, where there is no thinking there can be no delusion, hence it cannot be a delusion that thinking occurs. If it is thought that it does, then ipso facto it does. Descartes takes this to imply that “I think” or “je pense” or “cogito” is true every time it is thought. But where does the I or moi or ego come from here? It is the background assumption that everything is either a substance or a mode of some substance that leads Descartes to suppose that he can get two for the price of one here, a double conclusion, the existence of both thought and thinker, from a single observation.

For if his fundamental insight is that thinking occurs, then since thinking seems clearly not a substance in anyone’s sense, his background assumption leads to the conclusion that there must be a substance of which thinking is a mode: a thinking thing or chose qui pense or res cogitans. Critics such as the eighteenth century German physicist and wit Georg Lichtenberg (see his Scrapbooks K76) claimed Descartes has no right to say “I think” or “je pense” or “ich denke,” but only “it is thinking” or “il pense” or “es denkt,” parallel to “it is raining” or “il pleut” or “es regnet,” where the “it” or “il” or “es” is a dummy pronoun required by grammar but denoting nothing. Thinking may be full of delusions, and one of them may be that in addition to an on-going stream of thinking, there is a continuing subject, the I or moi or ich, doing the thinking. What cannot be a delusion is only this, a Lichtenbergian might claim, that thinking occurs.
1.2 The Scope of the Mental

It is against such background that we can formulate a provisional definition of “dualism,” subject to later emendation. We may begin from the picture often found in introductory textbook discussions of early modern philosophy. There dualism, represented by Descartes, may be presented as simultaneous rejection of two rival monist views, materialism and idealism, represented by Thomas Hobbes and George Berkeley. And it may seem clear enough how these two monist views ought to be understood, namely, materialism as holding that everything is physical, and idealism as holding that everything is mental.

Yet these formulations are arguably suboptimal. For in much of philosophy since the late the nineteenth century a great role has been played by a third realm of abstract entities, including the objects of mathematics, that are neither mental nor physical — though there is a school, modern nominalism, that denies their existence. The formulations just enunciated have the unfortunate feature of committing both materialism and idealism to nominalism, entangling their adherents in issues very far from philosophy of mind. It will be appropriate, therefore, to consider more limited formulations, on which materialism holds only that everything mental is physical — every mental property is a physical property, every mental event is a physical event, every mental fact is a physical fact, and so on — while idealism holds only that everything physical is mental, both postponing questions about abstracta.

Since properties, events, states, processes, and facts are often classified as abstract, we must be careful how we understand such phrases as “mental fact” or “physical fact.” What is meant is not a fact that is a mental or a physical entity, but rather something more like a fact about the mental or about the physical. And similarly for the other categories listed.

Now a universal affirmative saying that “All As are Bs” is to be understood
to hold vacuously if there are no As. Thus materialism as just formulated includes two versions, of which one is **eliminativism**, according to which there simply are no items properly called “mental,” and if it is thought that there are, that is a delusion. This view is often associated with the name of Daniel Dennett, whose works do contain many passages in which he seems to advocate it – several commentators have suggested that his book entitled *Consciousness Explained* (1991), in particular, should have been entitled *Consciousness Denied* — though they also contain passages in which he denies that he ever adhered to such a view, and I do not wish to get entangled in exegetical questions, where I would have little or nothing to add to what has already been said on this head by Galen Strawson (2018).

In an earlier draft of this work I gave the eliminativist position some fairly extended discussion, but I was advised by experts that this was unnecessary, since “no one” (which I take to mean, “no great number of prominent philosophers specializing in the area”) holds that sort of view any more. I had feared otherwise given what I had read browsing the blogosphere; and given the unusually high standing (for philosophical works) of Dennett’s books in the Amazon rankings; and given the regular appearance of a stream of books advocating what appear to be eliminativist positions to be one by one negatively reviewed in the *New York Review of Books* by Thomas Nagel and others (see Nagel 2017, Searle 1995); and given what I had heard in a talk I attended by a psychologist at my own university, who compared the belief that we are conscious to the delusions of a psychiatric patient who maintained that he had a squirrel inside his skull. I was happy to be corrected, and will as advised curtail any further discussion of eliminativism, which indeed has always seemed to me, as it has to many others, to be already sufficiently refuted by Descartes’ fundamental insight as recalled above.

The other version of materialism, surely more widely held, grants that there
are items properly called “mental,” but holds that these can all with equal propriety be called “physical.” The mental is subsumed under the physical on this view, which for want of a better term I will simply call “non-eliminativist.” Idealism would likewise have two contrasting forms. Dualism, as the rejection of both kinds of monism, will have two components, an anti-materialist, holding that there exist items that are mental but not physical, and an anti-idealist, holding that there exist items that are physical but not mental.

Now as recently as a hundred years or so ago idealism in one form or another was widely accepted. Its following has since so dwindled that today it is difficult to find any open adherent among professional philosophers (apart from proponents of idealistic interpretations of quantum mechanics deriving ultimately from Eugene Wigner, whom I will simply ignore, the interpretation of quantum mechanics being just too large a subject to try to tackle in the compass of a short work mainly on other topics). I personally have only ever met one avowed idealist. So idealism may as well be dismissed from further consideration here, making the anti-idealist side of dualism otiose. That leaves us for now with just materialism in its non-eliminativist version and the anti-materialist side of dualism, both granting that there are mental items, with the former affirming, and the latter denying, that these are in every case also physical.

On this account, the materialist’s thesis is the universal affirmative “All mental items are physical.” The dualist’s thesis is not its contrary, the universal negative “No mental items are physical,” but only its contradictory, the particular negative “Some mental items are not physical.” This consideration is important because the range of what is recognized as mental has expanded over the past centuries, and doubtless will continue to do so. That is part of what was meant when at the outset I described the mind-body problem as a “restricted and rather specialized part” of philosophy of mind. Materialism must hold across the board
that any newly-recognized sorts of mental items are just as physical as the old, while dualism may set aside the question of the status of the new, and rest its case entirely on an old list of examples such as might perhaps already have been recognized by Descartes, though there are differences at least in terminology and emphasis between him and even his closest present-day followers.

Descartes’ most general term for the mental is “thought,” or its French or Latin equivalents, but he understands the word more broadly in some directions than we do. Notably he includes sensation or perception as a kind of thought, though indeed it comes in only at the bottom of his list of kinds of thought at the beginning of his Meditation III (where it is tempting to conjecture that he is mentioning items in order from least to most dependent on embodiment, as he judges matters). At the top of his list come affirming, denying, and doubting, and these are perhaps closer to what we today in the first instance count as thinking.

Today, however, as soon as a philosopher mentions thinking, note will be taken of a key distinction between **occurrence** and **dispositional** thinking. If I say some former office-holder is thinking that he will be increasingly missed in the months and years ahead on the grounds that just now he is saying, if not aloud, then at least to himself, “I will be increasingly missed,” then I am speaking of occurrence thinking. If I say he thinks that he will be increasingly missed on the grounds that this is something he tends to say, though just now he is not saying anything even to himself, let alone aloud, but sleeping dreamlessly, then I am speaking of dispositional thinking. It is activities of occurrence thinking that Descartes is speaking of when he speaks of affirming, denying, and doubting.

Present-day dualists’ most general term for what they want to claim to be non-physical is “experience” or somewhat redundantly “conscious experience” or more often just **consciousness**. Now it is a commonplace that most of our language being designed for speaking about things in our shared and objective external
world, attempts to discuss our personal and subjective internal worlds may very often end in miscommunication. So one cannot just say, “Let us restrict our attention to conscious states,” and expect to be understood at once. Rather, one had best offer both an alternative characterization of the general class of states in question, and above all a representative list of instances.

The most widely accepted characterization of conscious states is surely that of Nagel (1974), cited in almost every exposition, which describes them as states “there is something it is like” to be in; Stevan Harnad would amend this to “there is something it feels like.” Here what it is like when there is something it is like is called the “phenomenal character” or “qualitative character” or “quale” (plural: “qualia”) of the conscious state. The instances that will be of most importance here are as follows.

First, moved from the bottom to the top of the list, there are sensory states, states of having this or that sensation, the most popular examples involving visual sensation, specifically the sensation of color, especially of the color red. As Descartes would have insisted, this includes cases of visual sensation when hallucinating or dreaming. Closely related are states of imagination and memory, but these for present purposes hardly require separate consideration. Second, there are states of affect, states of having this or that feeling, where the kind of feeling involved is that exemplified by feeling anxiety or feeling euphoric.

A much-discussed example of a conscious state not so far mentioned is that of being in pain, which generally involves both a sensation and also an affect. The sensation may be described as sharp or dull, continuous or intermittent, and referred to a particular part of the body. As Descartes would have insisted, the part in question need not be actually present, as in the case of phantom limb pain in amputees. The affect is an unpleasant one, called suffering from the sensation, at least if it is at all intense. The precise relationship between sensation and affect is a
matter of some controversy that need not be entered into at this point.

Third, though less important for us, will be conscious thought, occurring thinking that there is something it is like or feels like to be engaged in. By contrast, dispositional thinking is not a form of conscious experience. For in dreamless sleep there is no feeling, and hence there is nothing it feels like to be dreamlessly sleeping, nor anything it feels like to be in a state that one can properly be said to be in even when dreamlessly sleeping, which includes dispositional thinking. But is all occurring thinking conscious? Since the days of Descartes it has become increasingly accepted that there can be occurring unconscious thoughts.

Now we sometimes speak of “thinking” on the part of gadgets that are not conscious. For instance, if we ask why, when the fireplace is in use, the rest of the house gets cold, we may be told that since the thermostat was installed too near the hearth, whenever a fire is lit it thinks the house is too warm, and turns down the furnace. We all say this sort of thing, though few if any of us, I presume, believe that something like a thermostat really can think in any but the metaphorical sense of behaving in some salient respects like an agent who does.

The genus of postulated unconscious thought that raises more serious issues than the metaphorical thinking of thermostats includes as perhaps its most conspicuous species the unconscious thinking posited in psychoanalytic explanations of behavior. If the occurrence of such thinking is granted, materialism is committed to regarding it as physical. But dualism need not say anything about it, since sensation, affect, and conscious thinking may provide enough examples to establish the particular negative that not everything mental is physical.

And even conscious thinking may be lower down on the present-day dualist’s list than it was for Descartes, and with it those perceptual or emotional states consisting of sensations or affects as the case may be, together with a lot of thought-content. Examples might be seeing the traffic signal change to stop or
worrying that one will be late to the meeting, in contrast to seeing something red and shining or feeling free-floating anxiety, which are closer to being what might be termed “raw feels.”

Note the entire absence from the list as described so far of such intentional states or so-called propositional attitudes as belief or desire. (I did not mention Patricia and Paul Churchland alongside Dennett when discussing eliminativism, though that label is often applied to them, precisely because their kind is mainly a view about propositional attitudes.) There is something that it is like to be engaged in conscious thinking, but we will not here enter into any questions about what are called the “contents” of the thinking being engaged in, or questions about when what it is like to be thinking that \( p \) is or is not the same as what it is like to be thinking that \( q \) (an issue about which there has been much discussion, including an extended debate under the title “internalism versus externalism”).

Materialists are committed to regarding states of believing and desiring as physical; dualists can remain noncommittal about them or simply ignore them. Important though issues of intentionality may be in their proper place, dragging them into discussions of the mind-body problem, as is done especially by advocates of various so-called representationalist theories, is something mistrustful dualists may suspect to be a kind of misdirection of attention, tending to facilitate materialist legerdemain or prestidigitation to make conscious sensing vanish and conscious feeling disappear. So I will leave representationalism aside, along with any other issues that pertain to intentional states. (But one prominent representationalist, William Lycan, will be met with later in a different context.)

Now it is one thing to be having a red visual experience, and another to be saying to oneself “I’m having a red visual experience.” This contrast is the root of the opposition often drawn in the literature between first-order awareness and second-order awareness of awareness, or between “conscious” and “self-
conscious.” One case of the word “conscious” being differently understood by different writers, against which I have warned, is that some use the word for a second- rather than the first-order condition.

An example is provided by Julian Jaynes (1976) with his doctrine of what he calls the “bicameral mind.” According to Jaynes’s mind-boggling hypothesis, some three millennia or so ago people did not deliberate but rather, when they faced a crucial decision, would hear voices, coming from the right side of the brain but interpreted as divinities, telling them what to do; and what he calls “consciousness” arose only when the voices fell silent. On the usage of “conscious” adopted here, this is a nonsensical claim. If people were hearing voices, they were having auditory hallucinations, which are a kind of auditory sensation, which are a kind of conscious experience; and so they were already conscious. Jaynes has fallen into writing “conscious” where on the usage adopted here one would have to say “self-conscious.” If one uses “conscious” as Jaynes does, one will need another word for “consciousness” in the sense it which the term is used here; “sentience” is available for this purpose.

It will be convenient henceforth in this discussion to let a restriction of the term mental to the examples high up on the dualist’s list be tacitly understood, without denying, or rather, while emphasizing, that a broader use is appropriate elsewhere in what I have called the “large and growing field” of philosophy of mind. After adopting this convention, but only because it has been adopted, materialism and dualism may be described in language that sounds more symmetrical, one claiming that all and the other claiming that none of the mental is physical. These formulations hide the dialectical asymmetry that only the dualist and not the materialist has some freedom to choose what is to count as mental.

So much, then, by way of preliminary characterization of dualism and materialism, the latter nowadays often going by the alias “physicalism,” pending
modifications when we come to consider how modality was in the last century brought back into the discussion.

1.3 Psychophysical Correlation

In much debate between dualists and materialists, both sides have assumed a particular version of Descartes’ view that the brain is the part of the body of a living human being that is most closely connected with thinking. They have assumed what will here be called the correlation principle, the thesis that mental items are correlated with physical items, and specifically neural or nerve-related items, and even more specifically encephalic or brain-related items, by laws of nature. So long as no particular contentious conception what it is to be a law of nature is written into the principle, this thesis has enjoyed wide assent.

For much of the last century the issue between materialism and dualism came down, for many on both sides, to the question whether the correlation in question amounted to identity, the strongest form of correlation there is, or something weaker. For a considerable period the main form of materialism was a view called by David Armstrong (1968) the “central-state identity theory,” but which I will call neuralism, according to which each mental item is quite literally one and the very same thing as its neural correlate.

It cannot be claimed that the correlation principle is an established scientific result comparable to, say, the conservation of electrical charge. And what makes the principle plausible is by no means our having discovered the exact laws of correlation operative in many cases, or even any. Our present-day lay and scientific language may not yet even contain all the mind-related and brain-related terminology ultimately required for the formulation of any pertinent laws; indeed, it almost certainly doesn’t. Such plausibility as the principle has is the result of our having in various cases gotten what seem closer and closer approximations to the
ultimate law as we have learned more and more about the neuronal mechanisms connected with one phase of conscious experience or another.

The fact that no generally accepted exact law is known makes it impossible to give an uncontroversially real example of a correlation. For the sake of giving examples it is customary to pretend that we do after all know the law in one important case. The pretense is that we know the neural correlate of the state of being in pain to be the state of having nerve-fibers of the C-bundle that are firing. Neuralism is then, under this pretense, represented as maintaining that to be in pain is to have C-fibers that are firing.

This is only a pretense because, for one thing, other kinds of fibers are involved with pain (and sharp so-called first pains as contrasted with lasting second pains or aches are said to involve the firing of C-fibers less than that of Aδ-fibers), and for another thing, C-fibers are involved in sensations other than pain (and the firing of histamine-selective C-fibers is said to be involved less with pain than with itching). But because the example is only a pretense, no actual knowledge of what C-fibers are or do is required to follow its typical role in debate. Another equally over-simplified, pretend or toy, example would take the correlate of euphoria to be an abundance of endorphins in the brain.

While something like the correlation principle has been widely accepted, one does often hear tell of anomalous monists who are said to believe only in token-token and not in type-type identity: each individual mental event may be supposed to be quite literally one and the very same thing as some individual encephalic event, while the same emphatically denied to hold for kinds of events or occurrences as opposed to single events or occurrences. Donald Davidson (1970) is often cited as an example of one who thus rejected psychophysical laws of correlation of mental types with physical types. For reasons of space among others I will have to slight views that have comparatively few adherents, and anomalous
monism appears to be one of these. But a further reason for setting the Davidsonian view aside is that on closer inspection there is much less to Davidson's rejection of psychophysical laws than there seems to be at first glance.

Hardly any view could be less attractive that one that makes each mental token something physical, but in a lawless, haphazard way, as if my being angry yesterday was my having too much yellow bile while your being angry today is your having too little phlegm. Davidson’s view is not really of this kind. He allows that there may be laws of type-type correlation, but insists that they do not have the kind of exactness that the laws of fundamental physics possess, or are supposed by many philosophers to possess.

What is unclear to proponents of the correlation thesis is what this claim has to do with them. For the correlation they suppose to hold is with neurophysiology, not directly with fundamental physics in a Davidsonian sense. And it has repeatedly been pointed out in the literature that, though neurophysiology is integrated through molecular biology with chemistry and thence with physics, none of the other sciences mentioned is reducible to fundamental physics in a strong sense implying that its basic notions admit exact definitions in fundamental-physical terms.

The integration of chemistry with physics, for instance, through the development of the quantum theory of the chemical bond, offers a physical account of what is going on in what are called ionic and covalent bonding, but not in a way that permits either “ionic bond” or “covalent bond” to be given an exact definition in fundamental physical terms. Rather, the account reveals the extent to which such notions are inexact, though chemists go right on using the expressions “ionic” and “covalent.” The laws of such sciences as biology and chemistry, and what is most relevant here, neurophysiology, consequently do not possess the kind of exactness Davidsonianism attributes to those of fundamental physics. (Indeed, this is true
also for large tracts of actually existing physics.)

Even more significant is the fact that the items Davidson is concerned with and that he calls “mental” do not include states of visual or auditory or tactile sensation, or of pain or anxiety or euphoria that will be our principal examples in what follows. He is, rather, concerned with beliefs and desires, a very different case, and one I have already set aside. Moreover, Davidson is not concerned with our first-person ascription of conscious beliefs and desires to ourselves, but rather with third-person ascriptions of beliefs and desires to others, in an attempt to make their behavior intelligible to us. His arguments turn crucially on certain theses of his about what he calls “radical interpretation” and takes to be involved in thus seeking intelligibility for third persons. All this being so, it is hardly exaggerating to describe Davidsonianism as having almost nothing in common with the mind-body problem as it concerns us here, beyond a few scraps of terminology.

Anomalous monism dismissed, the correlation principle will henceforth be assumed, at least in the sense in which the correlation of being in pain with having C-fibers that are firing is being assumed, though this is a concession on the part of dualism that some dualists may someday want to reconsider. As I have said, it may be viewed as a more determinate and explicit version of the Cartesian focus on the brain as the body-part most closely connected with thought-processes. A wider separation between Cartesian and contemporary principles becomes apparent when we look more closely at the Cartesian mathematical orientation towards physics.

1.4 Physical Uniformity

Descartes was a much greater mathematician than he was a physicist. Today, Cartesian coordinates are taught to middle-school mathematics students the world over, while Cartesian vortices are remembered only by historians who are specialists in the arcana of seventeenth-century protoscience. In physics,
Descartes’ strength was in general strategy. He was a founder of the program of attempting to explain many physical phenomena in terms of a few mathematically-formulated laws. Moreover, he found one such law, governing the refraction of light, and more or less foresaw that there should be some kind of law of inertia and some kind of law of conservation. His weakness was in specifics. Failing where Isaac Newton was to succeed, he never got the right conservation law, mainly through lack of the Newtonian concept of mass.

There is a certain picture, or caricature, of this aspect of Descartes’ view that I am told derives largely from G. W. Leibniz and his criticism of Cartesianism among his contemporaries, a picture according to which even if he had had the concept of mass, Descartes could not have accepted basic principles of Newtonian mechanics such as the conservation of momentum. On this picture, the law of the conservation of momentum, the product of mass times directed velocity, cannot be admitted by a Cartesian, but at most a law of the conservation of the product of mass times undirected speed.

A Cartesian might grant that conservation of momentum may hold throughout the mineral and vegetable kingdoms, and among the so-called lower or non-human animals, and in such parts of the human body as the heart and liver and spleen and kidneys, and even in most of the brain — but not in the pineal gland. At least the direction if not the speed of the motion of the fluid supposed to fill this part of the brain can fail to be conserved, because the fluid is deflected by the action of the immaterial soul. On such a view, laws of nature governing physical systems may have exceptions in the special case where the system in question includes this one crucial part of the human brain, where the material body is supposed to be influenced by an aphysical spirit, and vice versa.

In contrast to this kind of view, it has come to be widely accepted that the physical brain, and indeed any other part of the physical cosmos, is immune to
interference from souls or other aphysical agencies. (To be sure, though widely accepted, this is not universally accepted and notoriously is rejected Wignerite quantum idealism; but that I have already dismissed.) Sporadic miracles cannot be proved not to occur, though the skeptic will insist that the burden of proof lies with those who hold that they do, not with those who hold that they don’t. But (quantum and other) miracles aside, it is widely held among philosophers today that there is no interference with the physical by the aphysical that is regular and systematic enough that it would have to be taken account of in a scientific psychology.

Today, not only is the correlation principle widely assumed, but a further **uniformity principle** seems to be almost equally widely, albeit perhaps less explicitly, assumed, according to which matter and energy are everywhere subject to the same physical laws, even in systems of which a human body or part of one is a component. And where the laws are statistical rather than deterministic, they apply inside and outside the body and brain with no change in the statistics. All present-day research in neurophysiology, in particular, seems to take something like this for granted.

The content of this principle may be made clearer by an analogy. Consider spherical bodies about the size of a ping-pong or golf ball, constrained to move on a frictionless horizontal track one end of which is arbitrarily designated “left,” the other “right.” Velocity in the case of such constrained motion may be understood to be simply speed with a plus sign in the case of motion from left to right, and a minus sign in the case of motion from right to left. Now if spherical bodies $x$ on the left and $y$ on the right are moving with velocities $a$ and $b$ respectively, and if $b < a$, the two will collide, and thereafter move apart with some velocities $c$ and $d$ for $x$ and $y$ respectively. This description is purely **kinematic**, involving only size, shape, speed, and direction, but no **dynamic** quantities such as mass or others defined in terms thereof, such as momentum or energy.
Allowing for idealization, some of the laws that govern such cases can themselves be stated in purely kinematic terms. For instance, if in-coming velocities \( a \) and \( b \) for a certain pair of spherical bodies are followed by out-going velocities \( c \) and \( d \), then in a repeat of the experiment, the same spherical bodies with in-coming velocities \( a + t \) and \( b + t \) will have out-going velocities of \( c + t \) and \( d + t \) (as per what is called “Galilean invariance”). But there are other laws that govern such cases that can only be stated by bringing in dynamic quantities. For the \( c \) and \( d \) following upon the same \( a \) and \( b \) will not be the same in the case of two ping-pong balls as in the case of one ping-pong ball and one golf ball. To predict \( c \) and \( d \) one must take into account also the masses of the bodies involved.

In this sense, a class of situations described in purely kinematic terms may be subject to laws of nature with only some and not all of the pertinent laws being stateable in purely kinematic terms. The uniformity principle in effect says that the relation between the physical and the mental is not like the relation between the kinematic and the dynamic in this respect. As has been mentioned, sensed color was excluded in the seventeenth century from the list of features of objects on which physical science was henceforth to focus. What the uniformity principle in effect says is that such features will never have to be brought back in in order to state the underlying laws governing aspects of situations describable in purely physical terms, in the way mass has to be brought in to explain the kinematically-describable behavior of colliding balls.

I have earlier stressed the commonplace negative point that biology, though today thoroughly integrated with physics, is not reducible to physics in any really strong sense; but even more important is the positive point presupposed by the integration of biology and physics, that it is now universally accepted that matter and energy are subject to the same physical laws even in systems of which a living organism or part of one is a component — at least if the organism is not conscious
or sentient, a case we postpone for a moment. The opposing view, vitalism, that
crude may be affected in the case of all kinds of living things from microbes on up
by a so-called life-force or élan vital, has been discredited among biologists for
well over a century. The uniformity principle goes a step further to disallow
exceptions for conscious or sentient beings.

Still less than the correlation principle can the uniformity principle be
claimed to be an established scientific result. It does, however, seem a reasonable
presumption in the face of the failure of decades of work in parapsychology,
which aimed to come up with reliably reproducible results demonstrating the
transcendence through the mental of ordinary limitations of the physical. Today,
parapsychology seems nearly as thoroughly discredited (in scientific circles: I am
not speaking of popular culture) as is vitalist parabiology. No credible case has
been made for, say, the ability to transmit something from one brain to another
without any physical signaling, or the ability by thinking about them to alter the
probabilities of random quantum events. This last was the object of extended
efforts at my own university’s now-defunct Engineering Anomalies Research
Laboratory. I share the sense of some of my colleagues that the extended operation
of this laboratory was an embarrassment to our school, and I will here take the
uniformity principle for granted.

Fifty years or so ago, when parapsychology had such distinguished
philosophical defenders as C. D. Broad, it stood in higher repute in philosophical
circles than it seems to do today, and there was genuine disagreement over its
status. Belief in so-called paranormal phenomena seems to be no longer for
philosophers today what William James called a “live option,” as it was for James
himself. Now so long as there was a division among philosophers on this issue,
that disagreement was probably the most important philosophical division over the
mind-body problem. It was not the division between dualism and antidualism: the
division line ran through the dualist side, between dualists who rejected and
dualists who accepted something like the uniformity principle.

My formulations thus far have involved the notion law of nature, but many
would prefer a formulation in terms of the notion of cause, so let me add that the
division of interest among dualists more or less coincides with the division
between so-called interactionist dualism like that of Descartes in the seventeenth
century, for whom body acts on mind and mind acts on body, and the opposing but
still dualist epiphenomenalism of T. H. Huxley in the nineteenth century,
according to which body acts on mind, but not vice versa.

More than once I heard my late colleague David Lewis say that virtually all
present-day positions in philosophy of mind would have been considered forms of
materialism by Victorians. I conjecture that he was taking it for granted that
something like uniformity is now generally accepted, and anything like
interactionism therefore rejected, and that what he meant by his remark was that
for Victorians even epiphenomenalism, the surviving form of dualism, in denying
the influence in the physical cosmos of immaterial souls, was as good as, or as bad
as, materialist monism, outright denying the existence of souls.

Consider, for instance, the case of Huxley, known as “Darwin’s bulldog” for
his vigorous defense of the theory of evolution, coiner of the label “agnostic,” and
author of the quintessentially Victorian-era metaphor for epiphenomenalism,
comparing the relation of brain to mind with the relation of a steam locomotive to
its steam whistle, which striking as its shriek may be, still contributes nothing to
the motion of the machine. Epiphenomenalism is very far from the crude
materialist view of J. P. Cabanis that the brain secretes thought as the liver secretes
bile — a view implying that thought is, after all, a space-filling stuff or substance,
though a fluid that sloshes around the skull rather than one that collects in the gall
bladder. Yet Huxley seems to have been regarded as no better than a materialist by
his contemporaries, beginning with his opponent in a famous debate over the origin of species, Bishop Samuel “Soapy Sam” Wilberforce (so nicknamed from Benjamin Disraeli’s characterization of his manner as “unctuous, oleaginous, saponaceous”). And speaking of Victorians, was even Friedrich Engels a materialist in a stronger sense than maintaining that matter can produce thought, but not the reverse?

In henceforth taking the uniformity principle for granted, I am assuming the twenty-first century dualist views of interest stand on the opposite side of the line from where stood Broad, and before him James, and before him Descartes, from whom we may now take our leave. There remain under consideration, therefore, at this point only non-eliminativist materialism and epiphenomenalist dualism. We must next complicate this neat picture by making room for a third alternative, provided by what is known as “functionalism,” and in so doing we will see how modal considerations came back into the discussion of the mind-body problem in the later decades of the last century.
2 Functionalism and Supervenience

2.1 Machine-State Functionalism

 Debate on the mind-body problem was shaken up not long after the midpoint of the twentieth century by the suggestion of Hilary Putnam and others (beginning as early as Putnam 1960) of a new alternative to traditional idealism and dualism and materialism, known as machine-state functionalism or the “computational theory of mind.” It rests on an extended simile likening the operation of the brain to that of a computing machine, which is to say, to an inevitably imperfect approximate physical implementation of some ideal mathematical algorithm.

 The key feature of such machines on which the suggestion focused was this, that a stage of an abstract algorithmic process, such as carrying one from the tens to the hundreds column in performing addition, might be implemented by different concrete physical processes in different machines, perhaps the turning of a certain gear in the case of a mechanical adder and the transmission of an electronic impulse between certain locations in the case of an electronic adder. As is said, the same role of one-carrying may be realized quite differently in different machines, involving gear mechanisms in one, and electrical impulses in another.

 In the present context what will be crucial is less this multiple realizability, as it is called, than simply the conceptual distinction between the implemented and its implementation. Even if we are considering only just one single implementation of the operation of addition, say the ringing up of some customer’s grocery bill on an old-fashioned mechanical cash register, there still nonetheless remains an important distinction between the role or abstract software process of one-carrying and its realization by the concrete hardware process of gear-turning. The difference is as great as the difference, even if we are speaking of just one single performance, between the parts of Mimi and Rodolfo and the soprano and tenor who perform them.
The novel suggestion in the 1960s was that being in a given mental state might be identified with the brain’s being engaged in a certain computational process, which is to say, with its being at a certain stage in the execution of a certain program, approximately and imperfectly implemented by a certain physical process going on in the neural wetware, as some would call it. The mental state is not, however, identified with the wetware state, as a neuralist might identify it. The mental state is not supposed to be the realizing physical state itself, but rather, the role which that state realizes.

But if it is therefore not a physical state, still less is it a mental state as traditionally conceived by dualists. The view being sketched is, on the one hand, one that does not comfortably fit the letter of materialism as provisionally defined earlier, though it may perhaps be viewed as in some sense fitting the spirit. It is also, on the other hand, one that does not comfortably fit the spirit of dualism as earlier defined, though it may perhaps be viewed as in some sense fitting the letter. What we are considering was a genuinely novel alternative.

It is also a rather problematic one, and was quickly met by objections from Saul Kripke, though right down to the time of this writing these are still available in print only in second-hand accounts (notably Buechner 2012). Not to put too fine a point on it, the objection is that it is meaningless to speak of the computational state of a physical object, mechanical or electronic or biological, except relative to some chosen way of viewing the object as the imperfect concrete implementation of some ideal abstract program. We do not notice this relativity in the case the mechanical or electric adder, because these are artefacts created by a designer with the intention of realizing a certain program, and we naturally understand the question of what computational state the device is in relative to the designer’s intention.

In the case of the human body, brain included, even if there is a superhuman
designer, still for present purposes there might as well not be, since we have not been made privy to that designer’s intentions, and those intentions therefore cannot provide us with an answer to the question of how to view brains and their states as implementations of algorithms and their stages. Then implicit behind machine-state functionalism seems to be the expectation that some unique preferred way of viewing the brain as a computer will nonetheless eventually emerge, allowing mental states to be identified with algorithmic states relative to that favored view. This is a large and questionable assumption.

Objections were also put forward by Ned Block (1978, the “China brain” argument) and John Searle (1980, the “Chinese room” argument). Putnam himself developed a bundle of further objections to his own view, and eventually renounced it (by the time of Putnam 1988 if not sooner). Machine-state functionalism does not seem to have any large, active school of defenders among philosophers of mind today. But the “functionalist” label, along with some of the features of the machine-state view, have been inherited by other views. We will return to these later forms of functionalism shortly, but before doing so let me try to clarify just what the machine-state functionalist’s claim about the relation of the mental to the physical comes down to, since that relation will be similarly understood by later forms of functionalism, and that understanding is what will in this study be most important about functionalism as a position on the mind-body problem.

2.2 Supervenience and Zombies

Machine-state functionalism represents an early instance of a position implying that though it may be that mental items are not themselves physical, they are at least what is called “supervenient” on the physical. The key concept of supervenience is a modal one, concerned with possibility and necessity, and the
invocation of this notion was one of the principal ways in which modality re-entered discussions of the mind-body problem. The first appearances of the notion in philosophy of mind were in connection with Davidson and responses to his work, but though Davidson’s anomalous monism has been dismissed here, the influence of the supervenience idea transcended its context of origin, notably through the work of Jaegwon Kim (culminating in Kim 2005).

Here As are said to be supervenient on Bs if no A could possibly have been other than it actually is without some B being other than as it actually is; or equivalently, if in order for any A to have been other than as it is, some B would necessarily have to have been other than as it actually is. Since the adder could not have failed to be carrying a one without its gear having failed to turn, in the case of the mechanical device, or its impulse having failed to be transmitted, in the case of the electronic device, the computational state of the adder is supervenient on its physical state, be that mechanical or electronic. In a slogan, “role supervenes on realizer,” though the two are conceptually as distinct from each other as Hamlet and Olivier.

Analogously, according to machine-state functionalism, mental states or processes, being computational states or processes, supervene on the physical states or processes in the brain that are their realizers, without being identical to them. Claiming supervenience of the mental on the physical in principle means just maintaining that the mental could not have been other than it is without the physical having been other that it is; but in practice, what machine-state functionalists maintain is something more specific: not just that the mental collectively supervenes on the physical collectively, but that each individual mental item supervenes on some individual physical item, its neural correlate.

Note that if all As are Bs, then As automatically and trivially supervene on Bs. In particular, if everything mental is also physical, then nothing mental could
have been different without something physical having been different, namely itself. Thus identity implies supervenience, amounting to the most extreme case of supervenience, as it amounts to the most extreme case of correlation. Supervenience of the mental on the physical is a point of commonality between neuralism and machine-state functionalism.

In practice, what deniers of supervenience believe is generally that without the physical being even in a minimal way different, the mental could have been different in the maximal way, by not existing at all. In taking leave of Descartes we have put aside the question whether there are ghosts, minds without bodies. Undeniably there are bodies without mind: corpses or cadavers. It is common for deniers of supervenience to maintain that though there actually are not, there potentially could have been living bodies without minds, bodies that did not just lie still, but that moved around and emitted sounds just as we do when we are pursuing our desires or expressing our beliefs.

Such are zombies in a philosophical, as opposed to folkloristic or horror-movie sense, a term popularized especially by David Chalmers (1996), for a notion that had made sporadic appearances in the earlier literature, but was brought to special prominence in his work. The deniers of supervenience typically hold that there could have been a whole zombie universe, physically just like the actual universe, but devoid of consciousness. In such a universe all the physical laws that actually hold would still hold, but all psychophysical correlation laws that actually hold would fail.

As with dualism, so with supervenience, one immediately faces issues of formulation, leading to what Lewis called an “unlovely proliferation” of notions of supervenience (and a chaos into which an effort to impose some order is made by McLaughlin and Bennett 2018). For instance, suppose there could have been ghosts, as many contemporary materialists, among them Lewis, have been
preparing to grant there could have been, even while insisting that there aren’t. And suppose these ghosts had been of a kind incapable of interfering with the rest of the universe, being mere watchers, so to speak. Then in one sense the mental would have been different, since there would have been mental activities, those of the watchers, going on that there actually are not; yet the physical would have been just as it is, since by hypothesis the watchers would have no influence on it.

Would this, then, amount to a counter-example to supervenience? It seems it would with the formulation under discussion so far. To avoid this consequence, in the literature reformulations have been proposed, bringing in an increasingly complicated apparatus of so-called possible worlds. A more direct way of avoiding the unwanted consequence is simply to recall and insist that we are only interested in the mental in the case of living human, not hypothetical ghostly, subjects. If the only difference from how things actually are were that there existed ghostly watchers, then the mental understood as the human mental would not after all have been different, any more than the physical.

With the introduction of the notion of supervenience, we have now not two but three positions to contemplate: first, the traditional materialist claim that all mental items are identical with physical items; second, the claim that though they may be distinct from physical items, still they are at least supervenient upon physical items; and third, a stronger-than-traditional dualist claim that mental items are not even supervenient upon physical items, let alone identical with such.

Early proponents of the intermediate view often emphasized that they were offering an alternative distinct both from traditional materialism and from traditional dualism. Nowadays it seems adherents of the intermediate view tend to think of “physicalist” as an honorific they would like applied to themselves. So let us apply the label to them by courtesy, whether or not it is not theirs by right. Then dualism, in a new sense, a strengthened one, denying supervenience as well as
identity, will stand in opposition to two forms of physicalism, weaker than 
materialism as understood in my discussion heretofore inasmuch and insofar as it 
includes not only a harder form claiming identity, but also a softer form claiming 
only supervenience.

That Cartesian and contemporary dualism both take a modal turn may seem 
mere historical coincidence, but the two can be amalgamated in an understanding 
of dualism as the thesis that there is no necessary connection between the mental 
and the physical. For by the most basic laws of modal logic, absence of a necessity 
of connection will be equivalent to presence of a possibility of disconnection, 
meaning the possible existence either of ghosts, minds disconnected from living 
bodies, or of zombies, living bodies disconnected from minds, or both. Let us 
accept this formulation as our official version of dualism henceforth.

2.3 Analytic and Psycho-Functionalism

Now we may come back to functionalism in its later manifestations, known 
as causal-role functionalism. According to a functionalism of that kind, a mental 
state is associated with a certain role, describable in broadly physical language, in 
terms of causes and effects, that a physical state may play in human life. Then 
being in the mental state is identified with being in a state that realizes this role. 
One of the attractions of the view for many philosophers has been that multiple 
realizability of the same role makes room for the same mental state’s being 
associated with very different physical states, biological or electronic or 
mechanical, in different types of beings; but this feature will be of little concern 
here, since the mental lives of animals, robots, extraterrestrials, or whatever is a 
topic we have set aside.

The formulation just given urgently calls for illustration by an example, but 
as with neural correlates, only toy examples are available. So let us pretend that a
state realizes the **pain-role** if and only if it (i) is caused by tissue damage and (ii) causes avoidance behavior. This simple-minded characterization is obviously grossly insufficient, but enumerating the ways in which it is so will be instructive as to what a better characterization would have to be like.

Three obvious difficulties with the toy example are as follows. First, the inputs (and out-puts) mentioned have not been characterized in purely physical terms. (For “damage” involves a value judgment: does tattooing damage or enhance?) Second, the one potential cause (or effect) that has been noted is obviously not the *only* one. (Just think of psychosomatic pains.) Third, the causes (and effects) that have been or could be cited are as we all know from experience not fully deterministic and operate only as tendencies. (Pain may not be felt from tissue damage owing to neuropathy or anesthesia.)

A deeper difficulty is that a mental state like that of being in pain does not just mediate all by itself between physical stimuli and physical responses, but does so together with other mental states with which it interacts: the inclination, when such-and-such has caused pain, to avoid such-and-such henceforth can to a degree be resisted and suppressed, and whether it seems worth suppressing will depend upon one’s other mental states, among which functionalists would certainly want to count beliefs and desires. (The fact that an activity is painful need not invariably lead to its being avoided, if it is believed to gainful, producing some effect that is desired.)

This last feature implies that mental states cannot be given causal-role characterizations singly, as we have been pretending in the case of pain, but only in interacting groups, considerably complicating matters. It is recognition of this complicating feature that distinguishes functionalist views from earlier **behaviorist** views, which tried to make do with an account of mental life involving only dispositions, or tendencies of stimuli like tissue damage to be followed by
responses like avoidance. In the **psycho-behaviorism**, as I will call it, of psychologists such as B. F. Skinner, ordinary language terms such as "pain," which seem to denote some kind of internal state mediating between stimulus and response, are simply ignored, relieving the theorist of any need to analyze what if anything they might mean. In the **analytic behaviorism**, as it is called, of philosophers such as Gilbert Ryle (1949), and according to some readings Ludwig Wittgenstein, it is suggested that a correct analysis of the meaning of such terms will show them to refer only to observable behavioral dispositions, not hypothetical internal states. I will not attempt any more precise description of these old behaviorist positions, because they have been largely abandoned as inadequate.

The successor to analytic behaviorism is the **analytic functionalism** of David Lewis, which itself came in two successive versions, an earlier that may be called the “first-order” or “realizer” version in the young Lewis (1966) and then a later that may be called the “second-order” or “role” version in the mature Lewis. The earlier version claims that an expression like “being in pain” *means* something like “being in whatever state it is that is caused by tissue damage and causes avoidance behavior.” This supposed discovery of linguistic analysis is then coupled with the supposed discovery of empirical research that the state that is caused by tissue damage and causes avoidance behavior is the firing of C-fibers, to yield the neuralist conclusion that being in pain is having C-fibers that are firing.

The later version claims that an expression like “being in pain” *means* something like “there being a state that is caused by tissue damage and causes avoidance behavior, and being in it.” This is a second-order state in the logicians’ sense that its specification requires quantification over states, namely, the existential quantification “there is a state.” It is really only this second version that resembles machine-state functionalism in making room for multiple realizability, and perhaps the “functionalist” label should be reserved for it. The young Lewis
offered only an argument for neuralism, while the mature Lewis offered an alternative to it.

Analytic functionalism of the kind I have been sketching is only one of two forms of causal-role functionalism, the other being what is known as **psycho-functionalism**. Perhaps the most important figure connected with this variant of the functionalist idea is Block, who was along with Jerry Fodor a founder of it (see Block and Fodor 1972), though he later became a critic (Block 1978), much as Putnam was first a founder, then a critic of machine-state functionalism. Psycho-functionalism opposes the analytic functionalist idea that a definition of the pain role can be given in ordinary language, with its discovery being a result of apriori linguistic analysis, upholding instead the idea that its description must bring in technical vocabulary of cognitive science, some perhaps not yet even developed, with its discovery being a result of empirical research. We will return to this view shortly, but first analytic functionalism requires a bit more discussion.

Dualists find the claim of analytic functionalism, that the very meaning of “pain” can be given by an account of pain’s causal role, to be preposterous, for the following sort of reason, among others. There is a dangerous medical condition called **congenital analgesia**, reported to be found in a substantial fraction of the population of a village of northern Sweden, and in scattered instances or clusters elsewhere. Those affected by the condition are unable to feel pain, and so can come to know of tissue damage only by seeing it. The causal-role analysis that the analytic functionalist takes to give the very meaning of “pain” would consist entirely of principles that could be fully understood by a subject with congenital analgesia. But to the dualist it is obvious that this leaves out the main thing, the thing that cannot be known to such a subject, and that is a crucial ingredient in the meaning of word “pain,” namely, what it is like or feels like to be in pain.

The congenital analgesiac stands to pain just as the blind man in the
anecdote or parable in John Locke’s *Essay Concerning Human Understanding* stood to scarlet. In Locke’s example, after a studious blind man claimed to have at last understood what the term “scarlet” signifies, he was asked to say what that is, and responded “It is like the sound of a trumpet.” Now scarlet is indeed like the sound of a trumpet: someone familiar with the colors but not their names might perhaps be able to guess which is called “scarlet” upon being told that it is the one most like the sound of a trumpet. And it is something of an achievement to have excogitated this characterization of the color from hints contained in books or passed on by friends, in the absence of any visual experience. Still, Locke takes it to be obvious that his character does not understand what “scarlet” signifies, and his whole story has the air of a somewhat cruel joke.

The dualist is likely to agree that it is preposterous to think that a blind person (if totally and not just legally blind, and blind from birth rather than from disease or accident later in life, in which case some memory of visual experience may be retained) could understand the meaning of the word “scarlet.” More generally, it is denied by the dualist that one could understand the meaning of the word for a state of sensation or affect or the like, be it “red” or “euphoria” or “pain” or whatever, without having experienced or felt what it is like to be in the relevant state. This is the only normal way to acquire knowledge of the meaning of such a word (though if the state of knowing the meaning of a word is one to which the correlation principle applies, then though the usual route to inducing it involves exposure to a pertinent experience, for all we know there may someday be other, high-tech routes involving direct manipulation of cerebral tissue).

In an earlier draft of this work I devoted an appreciable amount of space to further discussion of and objections to analytic functionalism, since it is the most important example of a position on the mind-body problem holding that a version of anti-dualism is an apriori and not an aposteriori truth — what Chalmers (2003)
calls “type A” as opposed to “type B” materialism. As in the case of eliminativism mentioned earlier, I found specialists commenting on that earlier draft confident that this view was defunct. The first time I mentioned such comments in a public lecture I was immediately told by a member of the audience, an Australian philosopher much influenced by Lewis, that he remained an analytic functionalist. Nonetheless, upon looking further into the matter I do find that even very sympathetic commentators on Lewis, such as Wolfgang Schwarz (2015), are much more willing to defend analytic functionalism as a theory of intentional states, which are not our concern here, than as a theory of sensations or affects. So I will say no more about the topic, except for a few occasional incidental remarks.

There remains, then, only the last version of functionalism, the one that appears to have the most adherents today, psycho-functionalism. I will not go over here, even in summary, what Block and others have had to say, pro and/or con, about it, but merely note one potential query about its right to be considered physicalistic. For characterizing some subset of the truths about the causes and effects of pain as constituting a specification of its role or function — the ingredient of the view that gives “functionalism” its name — may raise suspicions that teleology is being sneaked into psychology, after being banished from physics around the time of Descartes, and much later from biology, where it constituted a major obstacle to thorough integration with physics, by Darwinian evolutionary theory.

Psycho-functionalists are fond of citing an analogy between a term like “pain” and one like “heart,” claiming the latter is plainly a functional term, a heart being an organ whose function is to pump blood around the body, and plainly scientifically respectable term as well. But there is room for doubt, or at least for a distinction. There is no question that the function of the Jarvik artificial heart, in the strongest sense of “function” as end or goal or aim or purpose, is to pump
blood around the body. For it was created by a designer, Robert Jarvik, with precisely the intention that this is what it was to do. As for the natural human heart, so long as biologists or natural historians were content to appeal in scientific contexts to design and purposiveness in nature, a similar view could be taken, that pumping blood is not only something the heart does do, but also what it is for doing; though what with our not being privy to the designer’s intention, it was perhaps more in the nature of plausible conjecture than of confident belief that the heart was created for this purpose.

But when the presence of hearts in many organisms is explained without invoking designs or a designer, merely on the grounds that organisms with an efficient means of circulating fluids around their bodies are more likely to survive — and by “survive” let us understand “survive and produce offspring,” and by “offspring” let us understand “fertile offspring” — teleological language may seem out of place. Many popularizers of evolutionary topics will indeed use phrases like “designed by natural selection,” as if the tendency of traits conducive to survival to become more pervasive in a population were a kind of creator or demiurge, endowed with volition and foresight; but of course it is not, and such phraseology is not at all acceptable in any literal sense. Whether, and in what way, and to what extent, teleological language can legitimately be indulged in even in a science seeking integration with physics is a much-discussed issue in philosophy of biology. (See Allen and Neal, 2020.) Some would cite natural selection itself as providing a legitimation of a metaphorical employment of teleological vocabulary, taking the “function” of an organ to be simply whatever feature of it makes its presence conducive to survival.

Even if teleology is not involved, there remains the question of how to delimit physically the subset of of the set of all facts about causes and effects that is to define what is being called the “function” of pain or whatever. This problem
of characterizing such a subset of facts did not arise for analytic functionalism, for
which the decision about which cause-effect relations are to be part of the
characterization of pain is supposed to be a matter of discovering by analysis
which such relations are already part of that characterization according to the
ordinary-language meaning of “pain.” Rather than enter further into this
contentious issue, I will at this point simply waive any requirement that the subset
be specified in broadly physical terms, and grant that psycho-functionalist views
may be considered physicalist inasmuch and insofar as they aim at physically-
stateable in-put and out-put conditions to define the roles that interest them, and so
long as the realizers of these roles are physical states.

To sum up the discussion to this point, three views remain before us, one
version of dualism and two of physicalism: epiphenomenalism, psycho-
functionalism, and neuralism. Each has a distinctive view of the relationships
holding among the following trio of states a given subject X may be in at a given
time t:

(1) **Experiential:** X is feeling pain at t
(2) **Functional:** There is a state realizing the role of being caused by tissue
damage and causing avoidance behavior, and X is in that state at t
(3) **Neural:** X has C-fibers that are firing at t

We may take the three positions on the mind-body problem under consideration to
be agreed (in pretending) that the three states just described are coextensive in the
sense that either all of them or none of them hold of a given X at a given t. We may
also take them to be agreed that this coextensiveness is an a posteriori discovery.

As for differences, epiphenomenalism, as a form of dualism, holds that there
is no necessary connection between the experiential state and the other two: it is
neither identical with nor supervenient on either. Psycho-functionalism holds it to be identical with the functional state and hence supervenient on the neural state. Neuralism takes it to be identical with the neural state. For the most part I will be concerned below with the opposition between neuralism and epiphenomenalism, leaving it to the reader to think through, with only occasional asides from me, how far what I have to say about neuralism would apply mutatis mutandis to psycho-functionalism.

Along the way to the very short list of views remaining under consideration, I have successively set aside other positions on the mind-body problem with the briefest of mentions of the usual label for the view and the name of one of its adherents: eliminativism as in Dennett, idealism as in Berkeley, representationalism as in Lycan, anomalous monism as in Davidson, quantum idealism as in Wigner, interactionist dualism as in James, machine-state functionalism as in Putnam, pyscho-behaviorism as in Skinner, analytic behaviorism as in Ryle, analytic functionalism as in Lewis. The reader who consults Chalmers’ influential taxonomy of views (2003) will find there are yet other positions I have passed over without even a mention, including so-called panpsychist views favored by Chalmers himself, of which the more extreme versions attribute something like consciousness even to elementary particles. The actual views of all the figures I have just named were or are a good deal more subtle and sophisticated than the cartoon versions I have briefly described.

Some other once-influential figures in the last century, however, came out with claims so silly as to make it difficult today to credit that they were ever put forward in earnest. Perhaps the prize should go to the behaviorist J. B. Watson (1913), with his claim that thinking is simply subvocalization, or slight, silent movement of the organs of speech. (A close competitor would be the Wittgenstein disciple Norman Malcolm (1959), with his claim that it is a conceptual confusion
to suppose, as scientific sleep-researchers did and do, that there are definite times during sleep when dreaming occurs.) It is not easy to believe today that Watson’s view was once taken seriously enough that it was not admitted to be decisively refuted until Scott M. Smith (1947) subjected himself to the dangerous experiment of being temporarily paralyzed with curare and kept alive on an iron lung machine: he verified that he could still think even with paralysis of his vocal apparatus. Let this be a warning that views fashionable today may come to seem ridiculous tomorrow.

One reason for restricting attention to the trio of epiphenomenalism and neuralism and functionalism is that my primary aim will be to illustrate some logical points about the notion of supervenience. For this purpose, it seems enough to have a few views representative of different combinations of options in play, and juggling more than three balls would only be a distracting complication. The more immediate excuse I have been giving so far, each time I have set a view aside, has been a judgment, my own or that of experts who advised me, that the view being dismissed does not have any large number of adherents today. Since this last might said of epiphenomenalism, too, I had perhaps better next address criticisms of that view before proceeding further.

Let me just remark before I do so that if one does restrict attention to the trio of views I am focusing on, and applies the label “physicalist” to both the second and third views on the list, then the first claim that can be made for the importance of the supervenience issue is simply that it marks the boundary between dualism and physicalism. Indeed, in some quarters today (see Stoljar 2015) “physicalism” has been revisionistically redefined as the view, not that everything is physical, but as the view that everything supervenes on the physical. (No exception need then be made for mathematical objects and facts about them, if as is usually held these are necessary and could not have been other than as they are, since then the key
condition, that in order for them to have been different, something physical would have to have been different, holds vacuously.)

2.4 Other Minds

If the dualist can manage to refute supervenience, both forms of physicalism will fall at once. But it is surely overly optimistic or unrealistically ambitious to imagine that either side of the dualism versus physicalism debate could outright refute the other. For a striking feature of the mind-body literature is a strong tendency for arguments that to one side seem decisive and dispositive to seem to the other side feeble and flimsy. This is so to a degree that makes some suspect that allegiance to dualism or physicalism may be more a matter of personal temperament — “tough-minded” and “tender-minded” in Jamesian terms — than of anything else.

Physicalism is the dominant or more fashionable view, and so much so that it is often treated as the default view, as if the burden of proof were obviously on the other side. The result is that many physicalists offer only very casual and cursory arguments against dualism before getting down to the issue that really interests them, that of which form of physicalism is the best (as when neuralists accuse functionalists of being too generous, and functionalists accuse neuralists of being too stingy, about acknowledging mental states in animals or robots or other non-human cases that have been set aside here).

In thus describing the situation I am largely paraphrasing Lycan (2009), a physicalist who is candid enough to acknowledge that there is nothing in the most common physicalist arguments against dualism that would need to be taken very seriously by a committed dualist. For a list of such arguments he sketches what a dualist reply might be. Unfortunately for present purposes, the replies Lycan suggests in some cases presuppose an interactionist rather than an
epiphenomenalist version of dualism.

But treatment of arguments specifically against epiphenomenalism is available in a thorough survey article by William Robinson (2015), who canvasses a variety of objections, indicates in each case what epiphenomenalist responses to the objection might be, and usually goes on to consider rebuttals to these responses, rejoinders to those rebuttals, and perhaps several further zigs and zags of the dialectic, all with abundant references. He does not claim to settle the issue, but to me he does seem at least to establish that there is no short, simple refutation of epiphenomenalism.

There is no need to redo everything that Lycan and Robinson have already done — and certainly no reason to redo Robinson’s reply to arguments from a causal theory of knowledge, than which there is no more thoroughly discredited position in all of general epistemology — but as a sample of the sort of thing they might say, going a bit beyond what they actually do say, let me take note of one stock physicalist, and especially neuralist, objection against epiphenomenalism. The objection is that epiphenomenalism robs us of any hope of solving the mind-body problem’s cousin, the problem of other minds.

For epiphenomenalism seems to imply that one can never have evidence that mental states, as opposed to neural states, are present, at least not in any case other than one’s own. It may therefore be objected that it would be a violation of the rule not to accept beliefs without reasons in their favor to posit that other people have mental lives. For any supposed argument by analogy from one’s own case would be an incredibly weak induction from a single case. Thus one of our most basic beliefs, a virtual “postulate of sanity,” as it has been called, would be left without justification were epiphenomenalism accepted; or so the objector argues.

To this it may be replied first of all that the circumstance that a view would seem to imply that we lack justification for some basic beliefs is at most a reason
for hoping, and not a reason for believing, that the view is untrue. And the existence of a reason for hoping that is not a reason for believing is first and above all a reason for being especially careful about avoiding wishful thinking.

That aside, the epiphenomenalist may grant that indeed one should not adopt new beliefs without evidence, but insist that adopting new beliefs is not what is at issue here. The belief that people around me are not zombies (belief in other minds) is of a piece with the belief that the future will resemble the past (belief in the reliability of induction) and that there is a physical universe around me and not merely a play of my own sensations (belief in an external world). These beliefs are not doctrinal innovations being suddenly proposed just now.

Such beliefs arise early in childhood — so early that they cannot plausibly be regarded as conclusions inferred by any sort of intellectual argumentation whatsoever. Indeed, they are really less articulated beliefs than spontaneous reactions to experience: I spontaneously react to other people as if they had minds, in part by projecting myself imaginatively into their situations; I spontaneously form expectations of what will happen tomorrow on the basis of what is happening today and what happened yesterday; I spontaneously react to my sensations as if occasioned by things outside myself.

The question is whether I must give up these instinctive beliefs or spontaneous reactions if I cannot provide an after-the-fact justification and articulate positive reasons for them (as per what is called “epistemic foundationalism”), or whether I may retain them until someone articulates convincing reasons against them (as per what is called “epistemic conservatism”). Is positive evidence needed to justify retaining old beliefs, or only to justify changing beliefs? The physicalist who offers the argument under discussion seems to assume the former.

So did Descartes, whose project begins by discarding anything that seems
open to any kind of doubts however far-fetched, even those based on nightmare
fantasies about the evil machinations of demons, and then turns to the attempt to
build up a structure of belief from indubitable first principles by indubitable
inferential steps. An important difference between contemporary dualism and
Cartesian dualism that I have not heretofore emphasized is that present-day dualists
are not engaged in any such project (and so are free to take a conservationist rather
than a foundationalist line in general epistemology).

To say all this, however, is not to say that my instinctive reactions associating
mental states comparable to my own to bodies distinct from my own are beyond
criticism. Sane adults are generally agreed in ascribing mental states to other
human beings, modulo differences at the margins concerning early fetal stages or
terminal patients in a vegetative state, and setting aside strange things said by
eliminativist philosophers (who in any case often subsequently deny having said
them). But differences do appear if we consider the views of young children, and
especially if we suspend our restriction of attention to human mental life.

The tendency to view nonhuman entities as having feelings, sensations and
affects, and thoughts as well, quite as much as the similar tendency to view other
human beings as having inner lives, emerges at too young an age for it to be the
product of any sort of intellectual argumentation, and is in this sense spontaneous.
It is not, however, unprompted. It is cued in part simply by the look of the entities
in question, and whether they have something similar in gross pattern to a human
face, but the main prompts and cues are behavioral. There can be little doubt, from
an adult standpoint, that young children’s reactions, cued in this way, tend to
overgeneralize, to attribute feeling where none is really present, though perhaps
they undergeneralize in certain directions as well.

Childhood memories are notoriously unreliable, but for what it is worth, I
seem to recall, from around age four, my first encounter with an automatic door at
the grocery store. Such gadgets were novelties at the time, the early fifties, at least in the part of the world where I grew up. And I seem to remember being disappointed when disabused of the idea that the door could, in any real sense, be said to see people coming.

For the maturing child does get disabused of at least the grossest errors, and older children come to think as typical adults do, that the cat really does hurt if its tail is pulled, while the teddy bear, even if it is a fancy kind that can be said to “talk” in the sense of playing back recorded messages, really does not feel anything, whatever one pretends while indulging in make-believe. There seems also to be a phylogenetic tendency parallel to this ontogenetic tendency, a tendency over time, as culture and civilization develop, for early animist beliefs to be abandoned.

How far there are philosophically defensible reasons for these tendencies is, however, a large question, and an epiphenomenalist may even consider it a point in favor of, and not counting against, epiphenomenalism that it recognizes that such issues admit no quick and easy answers. There will be some occasion to return briefly to them later, but for the moment I will set them aside, having found no obvious fatal objection to epiphenomenalism in this direction. There is a more basic objection made by neuralists against epiphenomenalists that deserves more extended examination at about this point.
3 Epiphenomenalism and Causation

3.1 Historical Analogy

The most influential objection against epiphenomenalism has doubtless been the simplest: I mean the objection that that it is wildly counter-intuitive to claim that our conscious experience never affects our behavior.

This has been debated between neuralists and epiphenomenalists, but where do functionalists stand? Well, on the one hand, the neuralist objection might be thought to tell equally against functionalism, since on a functionalist as much as on any dualist view, a conscious experience is not identical with the correlated neural state that contributes to the causation of behavior. It is identified with a certain second-order state, the state of there being a physical state of a certain kind and the subject being in it, rather than with the first-order physical state of that kind that the subject is in; and it is the latter, not the former, that serves as a cause.

Yet on the other hand, many functionalists seem to think that only a slight change of focus or formulation, or perhaps a nuance of difference in the understanding of the terms “cause” and “effect,” may suffice to turn the neuralist objection into a non-sectarian physicalist objection that they could equally endorse. Such issues about functionalism I will avoid as distractions, to concentrate on the opposition of neuralism and epiphenomenalism.

As to that, the neuralist asks. “Can it be seriously maintained that seeing a traffic light turn red never causes us to move our feet so as apply the brake pedal?” For this is precisely the sort of thing that epiphenomenalism does maintain, and for some neuralists that is reason enough for not taking epiphenomenalism seriously. What can the epiphenomenalist say in response to the charge of counter-intuitiveness? Quite a bit, it turns out.

As a first line of response, the epiphenomenalism may cite history, which shows that intuition or common sense has been found grossly mistaken in the past,
and that we have learned to re-educate it. As a second line of response, the
epiphenomenalist may cite scientific psychology as producing ever more evidence
of the unreliability of intuitive or commonsense judgments precisely about
conscious experiences as causes of behavior. As a third line of response, the
epiphenomenalist may cite a distinguished philosophical tradition of questioning
reliance on the concept of cause beyond the realm of everyday practical affairs,
rejecting employment of the notion in philosophical analysis or hard science or
metaphysical speculation. As a fourth line of response, the dualist may say to the
neuralist “You, too!” when it comes to charges of violating common sense: the
epiphenomenalist may claim that the identification of the mental with the neural is
itself counter-intuitive and uncommonsensical, and indeed a conceptual confusion
of a kind the behaviorist Ryle would have called a “category mistake.”

Let me elaborate these lines of argument one by one in turn, beginning with
the first. The view that embarrassment causes blushing, to give one putative
example of the mental influencing the physical, may be as firmly entrenched in
naive, primitive, pre-scientific, folk belief as was once the belief that the
alternation of day and night is caused by the revolution of the sun about the earth,
and yet be just as wrong. So the epiphenomenalist may contend. The wrongness of
the ancient commonsense belief in geocentrism has by now been generally
recognized, and there is no reason why the wrongness of the commonsense belief
in mental causation should not be generally recognized in the future. So the
epiphenomenalist may add.

Note that recognition, as a matter of scientific and philosophical principle,
that it is the neural correlate of embarrassment, and not embarrassment itself, that
causes blushing need not lead to any change in the way we speak in everyday life,
any more than recognition, as a matter of scientific and philosophical principle,
that the earth goes around the sun and not vice versa led to any change in the way
we speak of sunrise and sunset. If Jill asks, “Why did John’s face go all red?” and Jack answers, “He was embarrassed by what Joan was telling him,” Jill will not respond with a lecture on how it is the neural correlate and not embarrassment itself that causes blushing, any more than, if Copernicus asked, “Why has it gotten so dark?” and a neighbor answered, “You’ve lost track of the time; it’s later in the day than you think and the sun has sunk below the horizon,” Copernicus would have responded with a lecture on how it is the earth turning on its axis and not the sun moving that brings the evening dusk.

We very often speak as if something were the case that some or all of us, after a greater or lesser amount of reflection, recognize is not. Where it is only a few of the more reflective among us who see that it is not, philosophers will speak of an error theory of the status of ordinary discourse, while where the error has come to be more widely recognized, they may speak of fictionalism; but it is in either case the same sort of thing, or at most, different stages in the same sort of process. Copernicus may have spoken with crossed fingers, deliberately concealing his real opinion for fear of the inquisitors, but as geocentrism gradually gave way to heliocentrism, talk of the rising and setting of the sun came to be quite generally regarded as no more than a figure of speech, ellipsis for talk of the apparent rising or setting of the sun. And as dead metaphors become in time secondary literal senses, and perhaps ultimately primary literal senses, it is not surprising to find online dictionaries today defining “sunset,” not as “the sinking of the sun below the horizon,” but rather as “the apparent sinking of the sun below the horizon.”

Common sense thus is capable of being educated. The epiphenomenalist is claiming no more than that such education is needed in a case where today even philosophers, under the influence of false doctrines such as neuralism, have mostly thus far failed to recognize the need, just as in the sixteenth century even astronomers, under the influence of false Aristotelian and Ptolemaic doctrines,
failed to recognize the need to revise our understanding of why day follows night.

One difference between the cases of mental causation and geocentrism, a
difference that only strengthens the epiphenomenalist’s case, is that Copernicus
was able, when writing for posthumous publication directed at philosophical and
scientific colleagues, to say what really is going on when the sun appears to rise or
sink in the east or west, whereas we do not as yet know how to describe in
physiological terms the exact neural correlate of the feeling embarrassment, which
correlate is supposed to be the real cause in the case of the blushing.

We are able to describe it only as “the-neural-correlate-of-embarrassment-
whatever-that-may-turn-out-to-be.” If we actually took to so describing it, and
constantly saying such things as “John’s face flushed as an effect of the neural
correlate of embarrassment,” it would be like saying such a thing as “Joan was
awe-struck by the beauty of sun’s apparent setting.” The fixed phrase, “neural
correlate of,” like the fixed word “apparent,” being invariant and an object of
tedious constant repetition, would most likely soon be left to be understood tacitly.
We would then be back to the ordinary way of speaking, with a mental reservation
or purpose of evasion, a tacit understanding that it is not to be taken quite literally.
And then eventually it could be expected to follow that the meaning of the words
we use would shift to agree with what we really intend when we use them, as has
happened in the case of the dictionary definitions of “sunset” that are now current,
which take it to refer to a mere appearance of motion rather than a genuine motion.

Against all this it may be urged that to speak of the epiphenomenalist
position as going against common sense is to understate the case. The proposition
that none of our sensations or affects ever issue in vocalizations and bodily
motions, but only ride along atop the physiological we-know-not-what that really
causes our lips and tongues and our arms and legs to move, is not just a departure
from naive ways of thinking, like the proposition that it is the sun that stands still
and the earth that moves. It is, rather, a way of thinking of ourselves that may justly be called “intolerable.”

The epiphenomenalist position may be termed “incredible” in the literal sense that it is so intolerable a thing to try to believe that no one can succeed in believing it seriously and sustainedly, 24-7. One may profess it in philosophical debate, but will fall back into the ordinary way of thinking when one lets down one’s guard. This may be less a sober argument than an emotional protest, a *cri de cœur*, but in any case it deserves a reply. I think the epiphenomenalist’s best reply would be that, even allowing the foregoing opposing claim to pass for the sake of argument, it does not constitute a disproof of epiphenomenalism. Here a well-known aphorism of Nietzsche’s (item #121 in the third book of *The Gay Science*) springs to memory. In opposition to the claim that no one could go on living without postulating this or that, it is answered, “Life is no argument: error might be among the conditions of life.”

This epiphenomenalist reply might be strengthened if it were possible to offer an analogy, an example of a proposition having the features of being literally incredible in the sense that no one can believe it all the time, and yet being true. Any discussion of the putative example will be a digression, and there can be no question, therefore, of going into the matter at any length or in any depth. Yet I will hazard a candidate: what I will call **strong free will.** By strong free will I mean free will in a sense that is strongly a matter of freedom, incompatible with determination by chains of antecedent causes ultimately going back outside the agent, and equally strongly a matter of will, incompatible with occurrence by mere random happenstance; and indeed I mean free will in a sense that is incompatible with determination by a combination of antecedent causes plus random happenstance. Let me attempt, in the limited space of a digression, not so much to defend the claim that the denial of the existence of strong free will is an incredible
truth, as simply to explain what I mean by claiming that it is, leaving evaluation to the reader.

On the one hand, as to the denial being true, as to there being no such thing as strong free will, we do in fact seem often to think of the behavior of third parties as if it exhibited nothing of the sort. We seem to do so when we make forecasts about the probabilities of various reactions on the part of masses of people to proposed public-policy measures, for instance. And anyone who has read attempts by philosophers to explain the concept of strong free will, which in the literature goes by the label “agent causation” — who for instance reads commentators struggling to explain Immanuel Kant’s conception of transcendental freedom behind empirical determinism — may very well soon be led to doubt that the notion of strong free will is even genuinely intelligible.

On the other hand, it may I think with some justice be said to be intolerable emotionally and infeasible psychologically to think of ourselves — and our friends and loved ones when we are relating to one of them as an I to a you — as mere concourses of atoms, whose motions and vocalizations are the result of a combination of causal chains beginning long before the birth of anyone alive today plus mere random quantum fluctuations. What may seem unremarkable in third person cases may feel unthinkable in the second- and above all first-person cases. Something like this is one of the themes of the celebrated paper of P. F. Strawson (1962) on free will. My suggestion would be that we inevitably keep falling back into thinking of ourselves as having strong free will, that in a sense we do so every time we engage in practical deliberation, and ask ourselves, “What to do?”

For it is, I would claim, difficult to the point of being infeasible to keep in focus simultaneously the question what to do and the thought that what is going to be done is already determined by causal chains going back four hundred years, or is going to rest on something as haphazard as whether a cosmic ray will chance to
strike the brain at a certain angle during the next hour. And I think a well-known puzzle goes some way towards demonstrating the difficulty of my holding on simultaneously to the question what to do and the thought that it is already determined what I am going to do, though whether an analogous puzzle could be developed showing the difficulty of holding on simultaneously to the question what to do and the thought that what is going to be done will occur simply at random I will leave to the reader.

The puzzle I am alluding to is that known as Newcomb’s problem, though admittedly in that puzzle the additional perhaps problematic assumption is made that what is determined not only can be deduced by a superhuman intelligence from a description of the prior state of the universe together with deterministic laws, but further can be deduced in real time, as is said, deduced ahead of its actual occurrence, as a before-the-fact prediction rather than an after-the-fact explanation.

In one version of the Newcomb example, a superhuman being puts before me two boxes, one opaque, the other transparent and containing a thousand dollars, and offers me the choice between taking only the opaque box or taking both boxes, telling me that what has been put in the opaque box depends on its prediction of what I will choose after the terms are completely specified to me. The specification is completed by telling me that if the prediction is that I will take one box, a million dollars has been put into the opaque box, and if the prediction is that I will take two boxes, the opaque box has been left empty (or in a variant, has had only a cabbage put into it).

Arguments can be put forward in favor of choosing one box, on the grounds that, supposing it to be part of the conditions of the puzzle that I do completely believe in the being’s predictive powers and promises, I must expect that I will be left with a million dollars if I take only the one box, but only a thousand if I take both boxes. Arguments can be put forward in favor of choosing two boxes, on the
grounds that whatever is in the opaque box is already in it and can’t now be changed, and that whatever it is, I will be a thousand dollars richer if I take both boxes. While many philosophers regard the second argument as better than the first, and some regard the first as better than the second, I myself am inclined to see a genuine antinomy here, in a more or less Kantian sense, showing the incompatibility of the assumption of the puzzle, that what I am going to do is already determined, and presuppositions of my asking myself what is to be done.

Whatever one thinks of this suggested analogy, I would underscore, as already constituting a sufficient reply to the original objection, the Nietzschean point that a feeling that it is intolerable to think otherwise, or even the outright psychological infeasibility of seriously and sustainedly thinking otherwise, is no guarantee that what we are habitually inclined to think is true.

3.2 Scientific Evidence

But an epiphenomenalist might well wish to move beyond this first, purely defensive, line of response, and go on the offensive. Moving on, then, to the second contemplated line of response, the epiphenomenalist may cite scientific psychology, introspective and clinical and experimental alike, as indicating that common sense is demonstrably at best extremely unreliable when it judges that this or that conscious experience has played a causal role in generating this or that behavior.

The basic situation can be illustrated schematically by arrow diagrams of cause-and-effect relations, with causes to the left of and sometimes above their effects, connecting an external stimulus with a conscious sensation and its neural correlate as well as a behavioral response. There will be occasion to refer to a half-dozen such diagrams, as follows.
(1) stimulus \(\rightarrow\) sensation \(\rightarrow\) response

(2) stimulus \(\rightarrow\) correlate \(\rightarrow\) response

(3) stimulus \(\rightarrow\) correlate = sensation \(\rightarrow\) response

(4) stimulus \(\rightarrow\) correlate \(\rightarrow\) response \(\downarrow\) sensation

(5) stimulus \(\rightarrow\) intermediate \(\rightarrow\) response \(\downarrow\) correlate = sensation

(6) stimulus \(\rightarrow\) intermediate \(\rightarrow\) response \(\downarrow\) correlate \(\rightarrow\) sensation

In many cases, common sense would posit pattern a pattern like (1), whereas the uniformity principle suggests a pattern like (2). Now neuralism would analyze the situation in terms of a pattern like (3), which is in agreement with both (1) and (2). By contrast epiphenomenalism would have to analyze the situation in terms of a pattern like (4), which though it agrees with (2), disagrees with (1), the commonsense picture. Such is the neuralist’s causal argument, reduced to pictures: neuralism can agree with common sense, dualism must take common sense to be making the error of confusing a sequence of two successive or anyhow separate effects, the sensation and the response, of a common cause, the correlate, for a sequence of cause and effect. This is not an uncommon type of error generally speaking, but it is one unwelcome to have to recognize in this particular case.

The epiphenomenalist line I will be sketching would say that in many cases scientific results indicate that the real pattern must be one that the neuralist and
epiphenomenalist views would have to analyze in terms not of patterns (3) and (4), but rather of patterns making room for an intermediate as the true cause of the response. The neuralist version would look like (5), whereas the epiphenomenalist version would look like (6). Since both (5) and (6) are equally incompatible with (1), the supposed advantage of neuralism, that it agrees with common sense, where epiphenomenalism differs, disappears. Both views must say that common sense is in error, though the two views give different accounts of just what the error amounts to or consists in.

To begin with a simple instance of the situation depicted in the diagrams, absolutely untutored common sense may assume that when a person touches a hot stove, the resultant pain causes that person to jerk away the finger, indeed the whole hand, an instance of the stimulus-sensation-reaction pattern with the stimulus being contact with the stove, the sensation being pain, and the response being jerking away the hand. I have even seen this untutored opinion asserted as fact in print by a respectable philosopher, though I will name no names. No more than introspection, if one engages in it while accidentally burning one’s finger, is required to show that this is wrong: the hand is jerked away before the pain is felt, and the temporal gap, though short, is noticeable, at least if one knows to expect it.

What is going on neurophysiologically in this case seems to be well understood by specialists, at least to a first approximation. The contact with the hot stove leads to jerking away the hand by a reflex arc through the spinal cord, whereas the pain is not felt until an impulse has traveled up the spinal cord to the brain, which takes some fraction of a second not too small be noticed. The pattern involves an intermediate, something going on in the spinal cord.

The prudent anti-epiphenomenalist may therefore at most hold that suffering from the pain is responsible, not for the immediate reaction of withdrawing the hand, but only for a subsequent attitude of wariness around hot stoves, in
accordance with the adage “once burnt, twice shy.” The epiphenomenalist will deny even that, citing the neural correlate of the suffering (which seems to be some electrochemical goings-on in the part of the brain known as the amygdala) and not the suffering itself, as the true cause; but nothing I have said so far bears on that issue, except indirectly by suggesting that the testimony of common sense against the epiphenomenalist claim is essentially worthless, given the gross unreliability of the witness.

I have so far mentioned only introspective evidence of the unreliability of commonsense judgments about mental causation, and considered only errors on the topic of pain. But clinical and experimental psychology suggest many more substantial and interesting cases of the same kind of errors on the part of common sense, about a variety of other topics. I will take space to discuss just one. A strange situation uncovered by neurology over the past decades that has generated much discussion among philosophers, so-called blindsight, is the case I mean.

Blindsight was first discovered in monkeys, who even after the destruction of the part of the visual cortex peculiar to mammals showed a residual ability to respond to visual stimuli, which ability was then verified to be present also in human beings who had suffered damage to the corresponding area of the brain, though they reported a complete absence of conscious visual sensation. In the human case, researchers have found limited ability to guess what is going on in what would be the visual field (or on one side of it in the case of patients whose lesions affect only input from one eye), and even ability to reach effectively for objects reportedly unseen.

One even more curious phenomenon has also been encountered. A sighted person crossing a lounge full of scattered furniture, if asked how he or she manages to do so without bumping into it, will doubtless answer, in accordance with common sense, that the sight of the furniture leads to its avoidance. The pattern
here is the simple one of stimulus-sensation-response, with the presence of the furniture as the stimulus, the sight of it the sensation, and avoidance manoeuvres the response. In support of this analysis, the fact that if the lights are turned out, or the subject blindfolded, stumbling into furniture promptly ensues, can be cited as evidence that vision is indeed involved.

Now the blind, too, as is well known, are to varying degrees able to navigate around obstacles in a similar way, an ability that has been under discussion by philosophers and psychologists at least since the time of Denis Diderot’s *Lettre sur les aveugles*. This ability is not, of course, affected by turning out the lights, clearly indicating that for them vision is not involved, but some other sensory channel. A dozen or more interpretations of what is going on were put forth, though today I take it that it is agreed that bat-like echo-location is at least a large part of what is involved, and that those who exhibit the most astonishing capabilities — the case of Daniel Kish is especially often cited — are those who have been able, owing to the plasticity of the brain, to devote parts of it that would ordinarily be used in processing input from the eyes to processing input from the ears instead. This need not mean that the conscious experience involved is auditory, like ordinary hearing; on the contrary, traditionally the presence of obstacles up ahead has been described as being felt as a kind of pressure on the face, whence the term “facial vision.” (I am told that the sighted can to a degree acquire this kind of obstacle sense through training, if their auditory acuity is within the normal range; my own degree of hearing impairment is such that I cannot confirm this first hand.)

The case of the blindsighted is different from the cases both of the sighted and of the ordinary blind. For instance, they report no conscious sensation, visual, auditory, or tactile, but they do need light. In one celebrated case, of which Youtube videos are available on-line as I write, Beatrice de Gelder (2008, 2011) and coworkers reported that a blindsighted individual TN, who had lost conscious
visual experience after a couple of strokes, navigated around obstacles in a lighted corridor while denying any kind of visual awareness of their presence. The preferred interpretation here seems to be, to a first approximation, that input from the eyes is processed through two different channels in the brain, one normally leading to conscious visual experience, but the other having also important functions, in particular it seems in navigation around obstacles.

Reflection on such cases must sap confidence in the commonsense beliefs about the causal role of conscious visual and other sensation in explaining behavior. For they suggest that in the case of the sighted person crossing the lounge, the commonsense explanation is, if not wholly incorrect, perhaps incomplete. There may be a component in the avoidance behavior that is caused by neural states on the path of processing other than the one that leads to conscious visual experience. Perhaps some substantial fraction of the subject’s evasive motions are initiated by some stage of neural processing prior to, or anyhow distinct from, that which leads to consciously seeing the furniture.

When the traffic light turns red as we are driving towards it, and we apply the brakes, can we be quite sure that the activation of the motor neurons responsible for the relevant contractions of muscles in the legs was initiated only after visual sensation of the color change had reached consciousness? Perhaps the structures involved in blindsight are not responsible, but may not some other intermediate neural process very shortly before and on the way to the correlate of conscious sensation be at work? It is hard to exclude the conjecture apriori, even in the absence of any positive evidence for it.

Epiphenomenalism, of course, maintains that quite generally, if nothing earlier than the neural correlate is the initiating cause, at any rate that correlate itself is something distinct from the sensation proper that suffices to get the body moving. The study of the limited range of abilities available to the blindsighted
does not directly support such a view, but it supports the view indirectly inasmuch and insofar as, in common with a good deal of other material to be found in the literature of clinical and experimental psychology, it undermines confidence in commonsense explanations, and impugns the reliability of the stories we are inclined to offer ourselves about why we act as we do.

3.3 Philosophical Critique

Having given the example of blindsight, I will not multiply examples, though others could be cited. Let me move on, instead, to the third line of response contemplated earlier, that of questioning the use of the concept of cause outside the realm of immediate practical affairs. I am generally leaving functionalists to speak for themselves in response to issues raised by the neuralist’s causal arguments, but I may just mention that functionalists doing so have sometimes been known to claim that it is not the cause-effect relationship in the strictest sense that is important, but some other, nearby relationship such as the one they call “counterfactual dependence.” The line I am suggesting for the epiphenomenalists at this point goes much further, not merely downplaying but, so to speak, denouncing the notion of causation. A line the epiphenomenalists might take, I am saying, would be the bold one of challenging the unquestioning and uncritical acceptance of the notion of causation, in terms of which the neuralist’s objection, and the two dualist responses considered so far have alike been couched. This third route seems, so far as the published literature is concerned, generally a road not taken. Yet I want to put it on record that the road is there, beckoning and inviting, for the epiphenomenalist courageous or quixotic enough to want to explore it.

Again the issue goes back to the days of Descartes, or almost. Descartes’ immediate successors were as unquestioning and uncritical in their acceptance of the notion of causation as he was himself, and yet the principles they found evident
and compelling were very different from his. For Descartes, if I stub my toe against a stone, this causes a motion in the fluid or so-called animal spirits supposed to fill the nerves, leading from the toe to the brain and ultimately the pineal gland, and the motion in this last key organ then in some unexplained way causes a sensation of pain in my mind.

For some of Descartes’ successors, however, advocates of the doctrine known as “occasionalism,” the events in my body only provide the occasion for God, the one and only true causal agent, to cause a sensation in my mind. Such disagreements as that between Descartes and these successors of his could only occur because, unlike the stone and the pain, causal connections are not perceived, but only inferred. This last observation was the entering wedge for David Hume in his critical approach to the whole topic of causation, most conveniently available in the seventh essay in his *Enquiry Concerning the Human Understanding*.

On Hume’s view the whole notion involves a kind of illusion or confusion. We experience a classes of cases that to us seem similar, call them A-cases, each followed by an experience of a case of another class that also seem to us similar, call them the B-cases. After repeated exposure to the temporal succession and constant conjunction of A-case followed by B-case, on next experiencing an A-case there arises in us a certain impression, a sentiment of anticipation, expecting another B-case, holding our breath waiting for the other shoe to drop. The idea we derive from this impression, of a connection between A-cases and B-cases, though it originates in ourselves, we project onto the world around us, and erroneously come to think of ourselves as having discovered out there a secret power within As to produce Bs.

In the preamble to the *Prolegomena to Any Future Metaphysics*, Kant defended Hume against some of his critics, those called “common sense philosophers,” noting that Hume never questioned whether the notion of cause was
right and even indispensable. And indeed, Hume did not suggest that we should
give up the notion, or even that we could give up the notion if we wanted to, for
everyday, practical purposes. We will, on his view, inevitably think in causal
terms. The issue, according to Kant, was over the origin of the notion, of which
Kant offered a very different and much more complex account than Hume’s, which
fortunately there will be no need to enter into here. In this connection Kant notes
that the issue of origin is closely connected with the issue of wider application of
the concept. And this is what was really concerned Hume.

For Kant’s remark is a delicate allusion to Hume’s criticism of the appeal to
the notion of cause outside everyday situations, a criticism that becomes most
prominent in Hume’s critique of the line of thought in natural theology famous as
the “argument from design” or “if you found a watch on a beach” argument, the
same argument that was to be called into question a century after Hume by
Darwinism. Hume’s critique is adumbrated in the eleventh essay in the Enquiry.
And to come to the crucial point, Kant, for all his differences with Hume, in the
end largely agreed with him about the limited applicability of the notion of cause
to topics that transcend experience. When it comes to arguments for the existence
of God as a First Cause, Kant is no more friendly to what he terms the
“cosmological argument” than was Hume to the argument from design.

A century and more after the death of Kant, discussion of the limitations of
the notion of causation was revived by Bertrand Russell at one stage in the
kaleidoscopic evolution of his views. He put forward sharply critical remarks on
the notion of cause in an essay (Russell 1913) where he characterizes it as “a relic
of a bygone age,” and indeed of the stone age. For Russell, it is not merely that
many of our beliefs about what causes what are naive, primitive, pre-scientific,
folk beliefs; it is that the very notion of “cause” itself is a naive, primitive, pre-
scientific, folk notion. Certainly it is not a notion of technical physics, and unlike
“energy” and “force” and “work” and “power” and “action,” it is not even a notion that has lent its name to some technical physical concept.

In response to Russell, Nancy Cartwright (1979), for one, has pointed to the continuing importance of the notion of cause, especially in decision-making. Without pretending to represent Cartwright’s views with any fidelity, let me just underscore this point. Suppose we would like there to be more Bs, and have observed that As are correlated with Bs, and are asking ourselves whether to undertake a course of action we expect would be followed by there being more As. What would happen if we did?

It is conceivable that what would follow would be an increase not only in the number of As but also in the number of Bs, maintaining the correlation observed so far. It is also conceivable that no increase in the number of Bs would follow, but only a weakening of the correlation that obtained before we acted. It is in just this sort of situation that we want to contrast “causal connection” with “mere correlation.” Whatever Russell may have intended, it was certainly not Hume’s intention to criticize the appeal to the notion of cause in such practical situations. For Hume the importance of skeptical doubts lies always in their checking our tendency to speculate beyond the realm of everyday affairs.

Perhaps the best-known latter-day response to Russell’s skepticism by a philosopher of physics is John D. Norton’s discussion (Norton 2003) of what he calls “causation as folk science.” After careful consideration, Norton ends with at best a very qualified verdict and partial rehabilitation. He observes that though there is no such thing as caloric fluid, and heat is not a material substance as was once thought, still in many ways heat behaves as if it were a material fluid, and as the work of J. B. Fourier and Sadi Carnot shows, it can be useful to think of heat in this way. He then suggests that it is the same with causality.

But Norton’s reconstruction of something like a notion of causation in
restricted areas of physics conspicuously fails, or rather, makes no attempt, to
reconstruct any general notion of cause and effect that would apply between the
mental and the physical in either direction, let alone unify all the sciences by
making the search for relations of cause and effect in their different domains the
central task of them all. A notion of causation that is or would be relevant to our
present discussion simply is not considered in what has perhaps been the most
influential qualified defense of what Russell dismissed as stone-age metaphysics.

And thus Hume, to a degree followed even by Kant, and with them Russell
in one phase or mood, with Norton only partly dissenting, form something of a
critical tradition, not concerned to advocate what is never going to come about, the
banning of causal talk from colloquial speech, but to advocate suspicion of the
notion in hard science and philosophical analysis and metaphysical speculation.
The focus in the critical tradition is not so much on rejection of commonsense
beliefs about what causes what, as with the first two epiphenomenalist replies to
the mental causation argument for neuralism discussed above, but rather on
rejection of extrapolation beyond everyday use of the notion, whether in natural
theology or in philosophical psychology.

I have already myself in discussing correlation and uniformity in my first
chapter started with the notion of law of nature rather than of cause in the central
role, rephrasing matters in terms of the latter notion only because it is so common
in the literature to do so. But the resources of the critical tradition have not yet
been much tapped by dualists.

3.4 Tu Quoque

A fourth and quite different epiphenomenalist line of response to the
physicalist’s argument from mental causation is also available, one that turns
around the neuralist’s accusation that the epiphenomenalist is departing from
commonsense. It is a response answering the claim that it is counterintuitive to deny mental causation of physical events with the counterclaim that it is far more counterintuitive to identify the mental with the physical. Physicalism, on this view, is not merely wrong but wrong-headed, a conceptual confusion, and this is something with which interactionist dualists might agree as much as epiphenomenalists.

The nature of the alleged confusion in neuralism is best indicated by analogy. And so I ask, what is the largest church in the world? An expert on religious denominations will doubtless cite Catholicism, an expert on ecclesiastical architecture, Saint Peter’s Basilica. Suppose a child were to overhear these answers and conclude “Catholicism is quite literally one and the very same thing as Saint Peter’s Basilica.” The conclusion is surely utter nonsense. Ultimately what I will suggest is that neuralism is nonsense of the same kind.

But if the conclusion that Catholicism is Saint Peter’s Basilica is nonsense, why is it nonsense? How could one argue that it is nonsense? The first answer that occurs to one may very well be phrased in modal terms, involving the hypothesis of something happening that hasn’t actually happened but presumably possibly might happen. A terrorist bomb might reduce the basilica to a pile of rubble; but if this happened when the basilica was empty, so that the Catholic clergy and laity from Pope Francis on down escaped, shaken a bit perhaps, but with their core beliefs and practices intact, then in that case Catholicism would have survived while Saint Peter’s would have been destroyed, showing the two to be indeed two, not one and the same.

But let us not be too quick here. Granting the possibility of the bomb scenario contemplated, the basilica reduced to rubble, pontiff and parishioners alike escaping more or less unchanged, just why would such an occurrence constitute the destruction of Saint Peter’s but the survival of Catholicism? I do not
see how this question can be answered without bringing in the categorization or classification of Catholicism and Saint Peter’s, their characterization each as a thing of a certain sort, with certain identity criteria, telling us where one thing of the given sort is to be counted as leaving off and perhaps another counted as beginning. To anyone familiar with the names, “Catholicism” denotes an institution, but “Saint Peter’s Basilica” an edifice. The identity criteria for institutions and for edifices, telling us how much one can be reformed or remodeled and still itself remain there, and not a replacement, may not be — indeed surely are anything but — entirely clear. Still, isn’t it at least clear enough that reduction to a pile of rubble is enough to make for the destruction of an edifice, while survival of all the people involved with beliefs and practices unchanged is enough to make for the continued existence of an institution?

But now comes the important point. Arguably, once we have these judgments about classification and criteria in place, the modal claim, and with it the bomb scenario, may become superfluous. Isn’t the fact that Catholicism and Saint Peter’s Basilica are things with very different identity criteria, and therefore very different sorts of things, already enough by itself to establish that they are very different things? If so, any assertion to the effect that they are quite literally one and the very same thing must be a piece of nonsense.

Now the neuralist holds that being in pain, as a type, is the same as having C-fibers that are firing, and that my being in pain now, as a token, is the same as my having C-fibers that are firing now. Let me concentrate on the tokens, leaving the reader to think through the case of the types. To the dualist, it seems clear that my being in pain now is my having a feeling now, and that one doesn’t understand what the word “pain” means unless one understands this, and also that my having C-fibers that are firing now is my undergoing an electrochemical process now, and that one doesn’t understand what the phrase “C-fiber firing” means unless one
understands this. These things are as clear as or clearer than that Catholicism is an institution and that Saint Peter’s is an edifice.

To the dualist, it seems clear that the identity criteria for tokens of being in pain involve who is having them, and when, and what it is like to have them, whether are they dull or sharp, steady or intermittent, and to what body-part, real or phantom, are they referred — but nothing about physical particles. Likewise, it seems clear that the identity criteria for tokens of undergoing C-fiber firings involve the motions of physical particles down to individual ions — but nothing about what anything feels like. Indeed, to the dualist these claims seem clearer than any claims one might make about the identity criteria for institutions or for edifices.

To some dualists, the conclusion may seem inevitable that since the identity criteria are radically different, the things are of radically different sorts, hence are radically different things. If so, the neuralist’s assertion that they are one and the same thing is absurd, like a childish confusion of “church” in the sense of religious denomination with “church” in the sense of house of worship.

The principle that different identity criteria make for different things will, to dualists who accept it, make short work of either physicalist view in philosophy of mind. For what has just been said against neuralism would apply a fortiori against the functionalist claim that to be in pain is to be in a second-order state, since though it may be obscure what are the identity criteria for being in a second-order state, it seems clear enough that they are if anything even farther away than are the identity criteria for undergoing electrochemical processes from the identity criteria for feeling sensations.

Now the dualist’s concession of the correlation principle seems to give the neuralist a dialectical advantage. The dualist will note that, on the neuralist’s own showing, the discovery that pain is correlated with C-fiber firing is aposteriori.
This means that before the correlation was established, we must have had some way of knowing in what situations pain was present, so that we could check what neural conditions were present in the same cases. Pain must have therefore have some feature by which we recognize its presence, which feature is not mentioned in the physical description of C-fiber firing. Its seems clear what this feature must be: feeling the way it does, and to begin with, being a conscious state. And this, the dualist may suggest, the neuralist’s account, or any physicalist account on offer, is leaving out.

But the neuralist will note that, by the dualist’s own concession, being conscious itself must have a neural correlate. Let us pretend for the sake of example that we know what it is, and that it involves the synchronized firing of neurons in the brain at a certain frequency (as per Crick 1995). The neuralist will suggest that consciousness is not merely correlated with but outright identical to such neural firing. If so, then a physicalist account of pain that brings in such firing will in one sense at least not be leaving the conscious nature of pain out of account after all, though it will be mentioning it not as consciousness but as synchronized nerve-firing. Quite generally, if the dualist mentions something mental the physicalist does not mention explicitly as such, and claims the physicalist is leaving it out, the physicalist will reply that it has not been left out, inasmuch as it is identical with something physical, namely, the neural correlate that even the dualists concedes there to exist.

Considerations involving classification and identity criteria, however, promise to cut through this dialectic, and directly establish that the correlations are not identities. Of course they only establish such a conclusion given certain assumptions. In the background is the assumption of the following four points.

(1) Things come in sorts.
(2) Sorts come with identity criteria.

(3) Things of the same sort count as identical when and only when the identity criteria are met.

(4) Things of different sorts never count as identical.

The import of these principles can be illustrated by simple examples. Ad (1), Comet Hale is of the sort *material body*; ditto Comet Bopp; but woad, the dyestuff produced from *Isatis pictoria*, is of the sort *material substance*; ditto indigo, produced from *Indigofera tinctoria*. Ad (2), the identity criterion for material bodies is *being at the same places at the same times*; that for material substances, *having the same composition*. Ad (3), comparison of and calculations from the coordinates provided by Alan Hale for the comet he observed with those provided by Thomas Bopp for the comet he observed, show that they were at the same place at the same time, and hence are the same material body (now catalogued as C/1995 O1). Similarly, chemical analysis shows that woad and indigo have the same composition, hence are the same substance (with gross chemical formula $C_{16}H_{10}N_2O_2$). Ad (4), it would be nonsense to claim that Comet Hale is woad, or that indigo is Comet Bopp, since this would be identifying items of distinct sorts, with distinct identity criteria.

To be sure, many cases of so-called scientific identifications discussed in the literature are *not* of the foregoing kind, where two items of the same sort are identified by application of established criteria. Neuralists have been known to cite among other cases the identification of heat with random molecular motion as a precedent for what they are proposing. What is going on in this case seems to many to be just this, that we first know heat by what kinds of things it does, while it remains open what sort of thing it is that does this.

Above all, heat causes heat sensations. (More precisely, it causes the
sensation in normal human beings in a normal state of health, barring among other things neuropathy and the influence of anesthetic drugs; and it is the primary, though not the exclusive, cause of such sensations, since hot sauce and heat rub cause them, too, without being hot or giving off heat.) It remained for some time an open question whether this unknown cause was a material substance, hypothetically called “caloric fluid,” or a physical process, the random molecular motion earlier mentioned. The latter view is now generally held (though to speak just of random molecular motion, ignoring radiant heat, is to give only a simplistic first approximation to a more complex body of current physical theory).

As mentioned in passing earlier, Lewis’s view before developing the mature form of analytic functionalism in its final form was, not to put too fine a point on it, a neuralist view, but defended by a novel argument very roughly as follows:

As linguistic analysis shows, pain is whatever it is that is caused by tissue damage and causes avoidance behavior.

As empirical research shows, what is caused by tissue damage and causes avoidance behavior is C-fiber firing

Therefore, pain is C-fiber firing.

Here, too, as in the case of heat, we start with something being known by what it does, and move on to a conclusion about what it is. But the first premise here has been rejected, on the grounds described earlier, namely, that pain is a feeling, and knowledge of the definition or meaning of “pain” requires the ability to have, or at least remember or imagine, what it is like to be feeling it. Still, even if pain is not an instance as the Ludovician argument suggests, heat at least remains plausibly regarded as an instance of a scientific identification that is not a case of applying
identity criteria to two items already classified as of the same sort.

What seems to be missing, however, is any convincing example where two items, each already characterized as being of a certain sort, but of two different sorts, are nonetheless eventually concluded by scientists to be quite literally one and the very same, and the dualist must view the neuralist as proposing that unprecedented kind of allegedly scientific identification.

Like almost any view in philosophy, the set of background assumptions (1)-(4) will be questioned by some. It can at least be said for those principles that they are invoked in areas far outside philosophy of mind, and cannot be dismissed as simply an *ad hoc* set of assumptions made up by epiphenomenalists for the purpose of countering neuralism. For though this book is not the place for any extended excursion into philosophy of mathematics, it may be mentioned that a notable example of appeal to such principles is found in Crispin Wright’s response (in Wright 1983) to the so-called Julius Caesar problem, the answer to Gottlob Frege’s puzzle about how we know that the number two is not the famous conqueror of Gaul. It is because the identity criteria for numbers, as set forth by Frege in works examined by Wright, and endorsed by Wright himself, are so very different from those for human beings, that no human being, not even one officially deified by the Roman Senate, can ever count as a number.

Unfortunately, the history of controversy over the Caesar problem is not especially encouraging about the prospects for eventually reaching a consensus on other issues that may resemble it. There is a danger debate will be reduced to flinging back and forth intuitions about proposed identity criteria for some key sort; or if not about the identity criteria for some sort of entity, then about the prior sortal classification of some key items; or if not about particular sortal classifications, then about grand general principles as the four listed above; or if not about such principles, then about the locus of the burden of proof. There is a
danger discussion will simply reach an impasse. And this danger brings me back to the supervenience thesis.

The first claim I made for the thesis was that it marks the boundary between dualism and physicalism. This is of importance if one thinks the dualist needs to reply to psycho-functionalism, a physicalist view distinct from neuralism, in addition to neuralism itself. But even if one takes the kinds of criticisms of functionalism put forward in the literature to be decisive, and takes it that neuralism remains the only viable physicalist view, the threat that debating directly about identity claims will reach a stalemate provides a reason for side-stepping the identity question and taking up the question of supervenience, the question whether certain correlations hold of necessity, rather than the question whether they amount to identity. The second claim that can be made for the modal turn involved in the supervenience thesis is therefore this, that it may provide a way circumvent an impasse. It remains to be seen whether it as a matter of actual fact does so. Some further clarification of the content of the supervenience claim is needed, however, before this latter question can be addressed.
4 Physicality and Phenomenality

4.1 Primary and Secondary Qualities

The supervenience thesis involves the notions of the mental and of the physical as well as modal notions. Something has been said about the scope of the mental in the first chapter of this study. Something must be said about the scope of the physical in this one, before turning to modality in the next. For it is a defect of many discussions of the mind-body problem that they go on about the alleged distinction or lack thereof between the mental and the physical without ever spelling out just what the word “physical” connotes, and my own account has suffered from this deficiency up to this point.

The only worse procedure than offering no account of the nature of the physical is to dispose of the issue quickly and cheaply by saying that the physical is what is studied in physics departments. To take that line is to hold an important philosophical distinction hostage to the convenience or the whims of academic administrators. Indeed, it is to do worse with the question than academic administrators typically do, since they tend to place their physics departments together with allied sciences such as chemistry, astronomy, and geology, in schools of physical science. But it is not too much better to take the physical to be what is studied in schools of physical science, since many branches of the life sciences, beginning with molecular biology and including the case of most interest in the present context, that of neurophysiology, are as mentioned earlier closely integrated with work done primarily in physical science departments.

The nature of this integration is a major topic in general philosophy of science, about which I will, for lack of space and of expertise, say little more here than I have said already in all-too-briefly discussing reduction and uniformity. I will simply take it for granted that the events and processes discussed in theories in the sciences just mentioned as they are conducted today may properly be called
“physical” events and processes, and that all the explanations in terms of such events and processes that are provided by these sciences may properly be called “physical” explanations. What requires more extended discussion is a restrictive condition, implicitly observed in all the relevant sciences, on what belongs to the theoretical domain of the science in question — a restriction that in the case of physics proper goes back to the time of Descartes, near the beginning of the period when natural science was separating from natural philosophy, and “physics” was ceasing to be the name of a branch of philosophical speculation, and becoming the name of a branch of scientific research.

For we must go back to a certain distinctions made in one way or another not only by Descartes but also by Locke and by scientists from Galileo Galilei to Robert Boyle, who all in one way or another distinguish certain features or qualities of things they considered in some sense more objective or real, from others they considered more subjective or apparent. The account of the principle of division varied from writer to writer, likewise the distinctive labels, if any, used for the first list and the second list of features, though primary versus secondary qualities won out in subsequent discussion.

The lists of examples of the two kinds varied as well, though generally size in the sense of length or area or volume, along with shape, or “figure” as it was called, would be mentioned at or near the top of the primary list, and color high up on the secondary. The features on the first list were those on which the new science was to focus, and they are such as to lend themselves rather directly to geometrical and more generally mathematical treatment. Making the division was important to the new, mathematicized direction physical science was to take over the centuries to come, a development mentioned early on in this book, but whose full significance we have yet to ponder.

No philosopher today can be quite happy with everything said by the various
seventeenth-century worthies in marking their division. Certainly we would wish
to say neither that the terms on the first list always refer to something objective,
nor that the terms on the second list always refer to something subjective. On the
contrary, the vocabulary and terminology both of shape and of color have a double
use. They are used both in describing our experiences even when not occasioned
by any external objects and in speaking of external objects of kinds capable of
producing such experiences whether or not they ever have done or will do so. Thus
the phosphenes seen as a result of rubbing the corner of a closed eye have both
shape and color, while galaxies are spiral or elliptical and contain red giant and
white dwarf stars whether or not they are ever observed by astronomers, and it was
so already long before there were any astronomers to observe them.

At most it might be argued that with a term like “size,” though it is indeed
used in connection both with objective reality and subjective appearance, and
though our access to objective size is in a sense through subjective size, as when
two measuring rods laid side by side are seen or felt to match, still the objective
use is primary, and the subjective secondary. Whereas, with a term like “color,” it
may be felt that the subjective comes first and the objective second: it is because
various frequencies of light appear red and green and blue to human subjects with
normal color vision that those labels “red” and “green” and “blue” are applied to
the frequencies in question; and the use of the color-terms for appearances long
antedates their use for frequencies, or the development of the undulatory theory of
light.

The most obvious difference between the two lists, though not every
seventeenth century figure focused on it, is the one hinted at in passing just now
when I wrote of measuring rods being “seen or felt to match,” namely, that items
like sizes or shapes appear through two sensory channels, visual and tactile, while
colors appear through only one. For this reason, the blind can understand “square”
or “rectangular” or “polygonal” in a way they cannot understand “scarlet” or “red” or “colored.” Ideally the latter were to play no role in physical theory, nor were any other qualities accessible only through some one, single sensory channel. And this was not a matter of some inadvertent oversight, but of a deliberate exclusion.

Needless to say, a great many more notions are dealt with today in physics and allied sciences than appeared on lists of primary qualities three or four centuries ago. I have already remarked how Descartes’ list was deficient in not going far enough beyond geometric and kinematic features to take in dynamical features. Of these a great many have been added to the basic notions of physics over the intervening years, from mass and charge to strangeness and charm.

But these seem to be all in the nature of features theoretically posited to explain empirical observations, rather than of qualities supposed to be themselves directly sensible. If so, their introduction into physical theory does not compromise the freedom of such theory from dependence on anything not fully intelligible to the blind. When these features are measured with instruments, such measurement generally reduces, allowing for some idealization, to comparison of primary qualities of those instruments, to comparisons of lengths or distances, namely, the distance from the zero to the unit mark on the instrument and the distance from the zero mark to the location of the pointer as the measurement is made.

Since processes of logical and mathematical definition are also independent of any sensory channel — and certainly of vision, as attested by the existence of blind mathematicians, already cited by Diderot — notions definable from accepted physical notions are likewise intelligible to the blind. To be sure, with parapsychology rejected and telepathy or extrasensory perception ruled out, to communicate anything at all, even pure mathematics, some sensory channel is needed. But there is no one channel such that it specifically is needed for understanding, and in particular vision is not needed.
It is common to contrast mere description with genuine explanation in science, but when the contrast is made, the latter is generally taken to include the former and then go beyond it to take in more also. If so, then whatever is physically explained must first of all be physically described or defined. The conclusion now looms that nothing not fully intelligible to the blind is physically explicable. The train of thought that has led to this conclusion obviously raises many large issue of general philosophy of science, far more than I have space or expertise to discuss. So I will claim no more for it than that the argument is heuristically suggestive.

It will be seen later that there exist other, related but distinct, arguments for the same conclusion in the literature, which have been widely influential. Let me just add that if there is anything to this heuristic argument, not only is anything physically explicable fully intelligible to the blind, but further this is something that might almost be called a “conceptual” truth, discoverable by analysis of what has been the project of physical science since secondary qualities were excluded around the time of Descartes.

What has just been said of vision and the blind applies equally to hearing and the deaf. Indeed, it applies even to the deaf-blind. Great determination and skilled assistance are both required to overcome the impediments presented by a lack of visual and auditory sensation, but it is no surprise to find Helen Keller, who famously had both requisites, telling us in her autobiography (1903: chapter 19) that she studied physics. And the only difficulty she complains of is that of getting items Brailled in a timely fashion. For physical science, even in the broadest sense, cannot tell us anything that cannot be told in Braille, nor teach us anything that cannot be taught to Keller. Indeed, that may be taken to be a hallmark of the physical in the broadest sense, the category embracing scientific theorizing in everything from physics proper to physiology and beyond.
4.2 Non-Physical Science

Perhaps the best way to understand the scope and limits of physical science is to compare it with such kinds of non-physical science as there are. I see three varieties. One is concerned with relations between conscious states and internal neural states. This includes work going on under the medical label of “neurology,” such as research on the neural bases of pain, and how pain might be reduced by suitably manipulating them. But it also includes work going on under in pure science, with no immediate medical application in view. It includes all work studying brain and consciousness, rather than just brain and behavior, or to put the matter another way, that part of the broader field of neuroscience that goes beyond purely physical neurophysiology. It is upon such work, seeking what are called “NCCs” (for “neuronal correlates of consciousness”) that the case for the correlation and uniformity principles rests.

Two more varieties of non-physical science remain, however, of which there has been no occasion to speak thus far. One is concerned with relations between conscious states and external physical stimuli. This includes some very old material such as work on projective geometry and perspective drawing concerning the relation between real and apparent size and shape. It also includes the oldest branch of experimental psychology, called “psychophysics.”

The other is concerned with relations between different conscious states, as represented by work in introspective psychology, a flourishing field a hundred years ago, which fell out of favor for a time, but in which work has picked up since (though not always under its old name, which may be associated with a specific approach that is far from universally shared). Also included in principle would presumably be work in at least parts of philosophical phenomenology, though that will have to be left aside here. Let me consider some key aspects of each of these two genera of non-physical science in turn.
To begin with psychophysics, it is explicitly concerned with discovering laws pertaining to specific sensory channels. The most basic result, applicable to several such channels, is Weber’s law, to the effect that the amount of increase in stimulus needed to produce a noticeable increase in sensation is proportional to the amount of stimulus already present: a bright light must be increased more than a dim one to be seen as having got brighter, and a heavy weight must be increased more than a light one to be felt has having got heavier.

Here we have two clear points of contrast with physical science. First, reported sensations are treated as data on a par with measured stimuli. Second, the results will not be fully intelligible to subjects lacking the relevant sensory channels. Keller, never having had the experience of seeing one light look brighter than another, or having forgotten any such experience she may have had in infancy before losing her sight, will not be able to understand fully what Weber’s law for brightness is even about, in the way we can understand what Weber’s law for brightness is about, or both we and she can understand what Weber’s law for heaviness is about.

During the period when the understanding of psychologists was darkened by the scientific ideology of behaviorism, Weber’s law was reinterpreted physically, as being about the minimal increase in stimulus needed, not to produce a noticeable increase in sensation, but to produce a different behavioral response. To begin with, the behavioral response was just that of reporting a noticeable increase in sensation. The behaviorizing of the law in this way seems at first simply a cheap trick or bad joke, since the only reason for interest in sensation-reporting-behavior is a belief that it is sincere and truthful and so tells us something about actual sensory experience. The behavioristic approach does, however, when pressed beyond human reporting-behavior, promise an extension of the law to creatures incapable of reporting, and perhaps even incapable of experiencing, sensation.
But much psychophysical science would entirely lose its point if subjected to behavioristic reinterpretation. Indeed, the theoretical core of the different branches of physical science always comes wrapped in psychophysical lore of this character. Thus in pure physics the core theory of black-body radiation concerns the mathematical relationship between temperature and frequency of emitted electromagnetic radiation, but this material is never taught to students without adding something not present in the mathematical formulation of the pure theory, without informing or reminding them, in the ordinary color vocabulary known since childhood, of what the various frequencies of light *look like*.

Equipped with this practical lore in addition to the pure theory, the physics or astronomy student presented with a distant hot body will be able to give a rough estimate of its temperature simply by seeing whether it looks black or red or yellow or white or blue. Students who lack normal color vision, and have to rely on a spectroscope, will be at a disadvantage, though perhaps not a very great one, since everyone will have to rely on instruments (and for that matter, calculations involving the Doppler effect, if the light source is moving) to get more than a rough estimate. Behavioristic reinterpretation would be pointless here. For what interests astronomers is how the stars will look to themselves, not how their behavior on peering into their telescopes will look to any rat psychologists who may have sneaked into the observatory to spy on them.

The wrapping in psychophysical practical lore becomes thicker with chemistry and geology and their flame tests and bead tests. All the senses are drawn in: not only do acids turn litmus paper red, but they taste sour. The importance of the wrapping for applying, and for that matter, for testing the theory should not be allowed, however, to obscure the physical character of the mathematically formulable theoretical core.

Turning to introspective psychology in a broad sense, the category of work
concerned with relations between different sensory states, special interest attaches to studies going back to the nineteenth century that led to the various so-called color solids. Consideration of these can help to clarify just what is and what is not knowable about color without actual experience of color vision, a topic given prominence in recent philosophy of mind especially by Frank Jackson.

In 2 Henry VI Humphrey Duke of Gloucester exposes in Act II, scene 1 the fraud of one Saunder Simcox, a knave claiming to have been born blind and to have suddenly and miraculously acquired eyesight through the intercession of Saint Alban. Pretending to doubt that Simcox can see well even after the supposed miracle, Gloucester asks him to name the colors of various articles of clothing worn by some of those present, which he successfully does. He then asks Simcox to name several of the persons present, which he cannot. The Duke then argues thus:

If thou hadst been born blind, thou mightest as well have known all our names as thus to name the several colours we do wear. Sight may distinguish of colours, but suddenly to nominate them all, it is impossible.

Let us call this Gloucester’s principle. Is it correct? Let us suppose that, rather like the character Mary in a famous example of Jackson’s (1982), who learns all there is to be learned about color and color vision while locked in a gray-scale room, Simcox had learned all that can be learned of color theory by a blind man, and retained it all in memory. Gloucester concedes that, upon receiving his sight, Simcox should be able to distinguish colors, and presumably this ability would extend to noting various relations among them, such as one being more similar to a second than to a third, or one looking like a mixture of a second and a third, and the like. Given color chips, and given enough time, he might by putting
similarly-colored chips next to similarly-colored chips, assemble a whole color solid. But Gloucester would doubt he could name any of the colors in it. Yet this conclusion is perhaps hasty if we make a generous or optimistic estimation of what Simcox can do with what he has learned, and without making use of information about what colors are shown by various objects that he knows by touch (for instance, that snow is white, or grass is green).

Among other things, he will have heard, while still blind, of a regular alternation of periods of daylight, when distinguishing objects by sight is easy, and periods of nighttime, when it is difficult, and more generally of a distinction between well-lit and ill-lit areas. Upon receiving his sight he will soon enough, probably without having to wait for nightfall, experience ill-lit areas; and then he will be able to distinguish one color at least on the color solid, since he will have heard of black as the color to which all things tend when illumination is poor.

Next white can be identified as the color most opposite, or least similar, to black, and the gray scale as the colors that look like mixtures of black with white, and the hues as the colors that seem to have no admixture of gray to them. These last, with most similar adjacent to most similar, form a ring, in which some are brighter, or more similar to white, and others are darker, or more similar to black. Most will look like mixtures of the colors on either side, but for four pure hues this will not be so, and yellow and blue can be recognized as the lightest and darkest, respectively, of these. It only remains to determine which direction around the ring from blue to yellow is the path through red and which through green.

If, as some suppose, the synesthetic perception of some colors as warm and others as cool is innate, and not the result of association of red and yellow with fire, and blue and green with lakes and seas, then red and green can be distinguished as warm and cool. If not, one may have to resort to counting the number of distinguishable shades, there being more between blue and green than
between blue and red, where lie the non-spectral purples. But with a half-dozen cardinal points on the solid identified, all other colors can be described as mixtures thereof in various proportions, and it seems all colors can be named, on the basis of information acquired while still blind. And so Gloucester’s principle may be called into doubt.

But note what our hypothetical Simcox could and could not come to know while still blind: everything about the structure, one might say, of the color solid or color space, nothing about the content; how each color is related to each other, but not how any of them looks absolutely. Study of the physical properties of the color chips, what frequencies of light they reflect and how well, and study of the physiology of the portion of the nervous system concerned with color vision, together might be able to explain — modulo the great unexplained fact of the correlation principle, elaborated with the assumption that similars are correlated with similars — how Simcox is able to make the distinctions and similarity and other judgments that he makes; and all this could be communicated in physical terms, in Braille. But what the colors look like, that could not have been communicated to Simcox while still blind, any more than the appearance of red could have been explained to Jackson’s Mary before she left her room and first saw red items for herself, or that of scarlet to Locke’s subject. The phenomenal character of the conscious experience of color lies outside the bounds of physical science.

4.3 A Lacuna in Explanation

For responses to standard physicalist objections to dualism in general and epiphenomenalism in particular I have referred the reader to the work of Lycan and of Robinson. There is, however, one such objection, the first on Robinson’s list, about which a little more may be said now that we have looked at the distinction
between physical and non-physical science. The objection in question, influential since the times of James onward, is an argument from natural selection that in effect claims that Huxley’s intellectual position, combining militant Darwinism with an early and forceful enunciation of epiphenomenalism, was incoherent. The objection exposes an important division among physicalists, one that seems to cut across the division between neuralists and psycho-functionalists.

The evolutionary objection is just this, that since natural selection concerns what traits of an organism are conducive to survival and the production of fertile offspring, and since survival and production of fertile offspring are purely physical matters, if mental states are assumed to have no physical effects, then natural selection cannot explain their presence. To illustrate the point, let us suspend for a moment our restriction of attention to human mental life.

Consider, then, two hypothetical species of butterflies, with overlapping ranges, very similar in appearance as caterpillars, but with very different patterns of bright coloration as adults. Since the bright colors must make them more conspicuous to predators, how could it be explained in evolutionary, adaptationist terms? One partial answer might be that the distinctive patterns enable the butterflies to recognize others of their same species, so that they do not waste any reproductive effort on cross-species unions that could at most produce sterile offspring like mules, if they produced any at all.

So far, so good. But what are the colorful insects seeing as they flutter about looking for mates? Setting aside various complications (their possession of more than three types of color receptors, their sensitivity to the polarization of light), we may take them to be, like other butterflies, sensitive to differences in frequencies in the orange to ultraviolet range, incapable of distinguishing red from black as we can, capable of distinguishing ultraviolet from black as we cannot.

One can imagine the creatures having their whole spectrum shifted relative
to ours, so that what to us looks orange to them looks red, what to us looks yellow
to them looks orange, and so on, with what to us looks violet to them looking blue,
and the ultraviolet, which to us looks black, looking to them like violet. Or one can
imagine them seeing orange and yellow and green and cyan and blue and violet as
we do, but seeing ultraviolet as red. Or perhaps they see it as some color we cannot
even imagine. Or perhaps, their nervous systems being so very different from ours,
they see all colors in ways we cannot imagine. Or perhaps they have no conscious
visual experience at all, any more than a Venus fly trap has tactile experience.

From the point of view of explanation in terms of natural selection, none of
this matters. It is enough that different wave lengths have different physical effects
on the lepidopterans' eyes, leading to different electrochemical activity in their
nervous systems, which may result in differences in behavior, including
reproductive behavior. Quite generally, adaptationist explanations come down to
physical explanations, and the evolutionary explaining away in physical terms of
apparent teleology, apparent design and purposiveness, in nature was a giant step
in the direction of the uniformity principle. The other side of the coin is that if the
butterflies do have conscious visual experience, including color experience, natural
selection cannot explain why they have it, or what it is like to be having it, or why
it is like that. Or at least, natural selection cannot explain these matters directly.

It could indirectly help to explain them if they are linked to some purely
physical matters that natural selection can explain, provided the pertinent linkage
itself can also be explained. The proviso is crucial, however. For mental states are
linked to some purely physical matters, their neural correlates, according to the
correlation principle we have been assuming. But where we would need an
explanation of the linkage, of the psychophysical laws of correlation, we find only
what Joseph Levine (1983) has called an explanatory gap. The gap was noted
long ago by Huxley (1866: 193), who famously compared it to lack of explanation
of the link between Aladdin's rubbing the magic lamp and the djinn appearing. In the century and a half since Huxley's time the gap has never been filled or bridged, so there is considerable plausibility simply on grounds of pessimistic induction to the conjecture that it will never be filled or bridged. The objector takes the fact that epiphenomenalism implies that evolution cannot explain the mental to be a point against epiphenomenalism, but the epiphenomenalist takes it to be a point against the objector's apparent assumption that natural selection can gaplessly explain anything and everything in the world of life.

Levine himself is a sympathizer with physicalism, but his physicalism is what I will call lacunist as opposed to denialist, a physicalism that admits the existence of an explanatory gap, as opposed the physicalism of others, notably including Dennett, who seem to be in denial about the limitations of physical explanation in general, and evolutionary adaptationist explanations in particular. (Analytic functionalism, by taking a functional analysis stated in purely physical terms to give the very meaning of a term like pain, of course implies that there is no explanatory gap, and so constitutes a form of denialism; but analytic functionalism has already been dismissed.)

I have said that so long as interactionism was a live option, the most important division line among philosophers over the mind-body problem was the line running through dualism, separating interactionism on one side from epiphenomenalism and physicalism on the other. If interactionism is out of the picture, then it seems to me the most important remaining division is still not the division between dualism and physicalism, but rather one that runs through the physicalist side, separating denialism on the one hand from lacunism and dualism on the other. Let us now look a little closer.

When in normal health and under normal lighting, place a red tomato on the left side of a table and a green pepper on the right. Looking at the vegetable on the
left side, ask yourself, “Why is the experience of seeing red like this, rather than like that?” at the last word turning to look at the vegetable on the right instead. Dualism and lacunism agree that as a matter of principle there can be no physical explanation here, no physical answer to the why question, and in particular no evolutionary biological answer.

It should not be necessary to say — yet I fear it perhaps is necessary to say — that to deny there is an evolutionary explanation is not to suggest that there is a some other kind of explanation, a “creationist” explanation. Descartes’ and Newton’s seventeenth- and eighteenth-century contemporaries among philosophers and scientists almost all professed to be theists or at least deists, and would presumably have endorsed the view that whatever laws of nature hold, they do so because the creator so willed. But I think we may credit them all with having the wit to see that a general salute to a universal divine providence is by itself not going to explain the particulars of any specific natural law. Even on the theistic hypothesis, we have no explanation of psychophysical correlations.

I would further claim that there are reasons of principle, going beyond mere pessimistic induction, for the expectation that there neither will nor can ever be a purely physical explanations in this domain. The line of thought leading to this negative conclusion is nearly explicit in what has been said already. On the one hand, a heuristic argument was sketched suggesting the conclusion that there is no physical explanation of anything not in principle intelligible to the blind. On the other hand, the lack of full understanding of color terms on the part of the blind has been a recurrent theme. And of course, what has been said about color and blindness might be said as well about pitch and deafness, pain and analgesia, euphoria and anhedonia, and so on.
4.4 Phenomenal Concepts

In the literature, the phrase “phenomenal concepts” is used, following Daniel Stoljar (2005), as a broad term to cover (concepts of) color and pitch and pain and euphoria and other sensations and affects that have furnished us with examples. Various writers have offered various arguments tending to the kind of conclusion reached at the end of the preceding section, that is to say, pro-lacunist or anti-denialist arguments, arguments to the conclusion that what it is like to be having visual experience of some color, or auditory experience of some pitch, or to feel pain or euphoria, is physically indefinable and inexplicable. But the arguments I am alluding to, as compared with the argument of the preceding section, start from the opposite end, so to speak: not from the side of the concept of physics as it has developed since the first exclusion of secondary qualities in the seventeenth century, but from the side of the phenomenal concepts. Each argument attributes to them some special feature that is supposed to explain their detachment from the physical.

The literature discussing such arguments pro and con, of which Andreas Elpidorou has prepared a useful bibliography (2015), is quite large, and the proponents of different arguments differ quite a bit among themselves in their concepts of concept, and their claims about exactly what special features of phenomenal concepts make for their physical indefinability and inexplicability. The details need not detain us here, but it may be useful to sketch an account of a toy example that at least exhibits the form of the kinds of claims that are made in less simplistic analyses in the literature.

I begin by recalling two widely-recognized linguistic phenomena. First, just as English has a category of tense, and special grammatical devices for locating what one is speaking of in time relative to the time of one’s speaking of it, devices whose use moreover is in English obligatory and not optional, so other languages
have a category of **evidential modality**, and special grammatical devices for indicating the nature of the grounds for one’s assertions, devices whose use is in those languages obligatory and not optional. It is as if in English it were forbidden to say “There are stink bugs in the area,” without adding some such little phrase as “I see” or “I smell” or “people report,” whichever is appropriate to the case. The distinction between what is known by personal experience and what is known by report is perhaps the most basic evidentially modal distinction.

Second, there is **conventional implicature**, a feature of the meanings of certain words on account of which using one rather than another conveys a certain suggestion, in a way that can make an utterance infelicitous if the wrong word is used, but in a dimension not directly connected with truth. Such is the use of “**but**” rather than “**and**” to convey a suggestion of contrast. Having to do as it does with a dimension other than truth, the suggestion in question does not go away if an affirmation is changed to a denial, or to a question.

Thus if “poor but honest” is offensive for suggesting, after all the notorious misdeeds of banksters and crony capitalists, that it is poverty rather than wealth that contrasts with honesty, it is equally so in saying, “Is she poor but honest? I say she is,” which conveys “To be poor and to be honest contrast; and she is both,” and in saying, “Is she poor but honest? I say she is not,” which conveys “To be poor and to be honest contrast, and she is not both.” Indeed, the offensive suggestion is conveyed even without making any assertion at all, but simply asking, “Is she poor but honest?”

Now we can imagine a language in which certain words carry evidentially modal conventional implicatures, a language English* with a word “**red***” that works like this: If I say, “The fire engine is red*,” I convey “The fire engine is red, and I know from personal visual experience what red looks like;” while if I say, “The fire engine is not red*,” I convey “The fire engine is not red, and I know from
personal visual experience what red looks like,” and similarly for interrogatives and imperatives. Presumably notions from physical science are too impersonal for any consistent set of premises formulated solely in terms of such notions to imply anything about what I do or do not know from personal visual experience. And so no such premises can imply any conclusion non-trivially involving the word “red*.” There can be no definitional, and presumably no other apriori, connection between physical concepts, even in the broadest sense, and redness*.

If someone were to claim that in actual English the actual word “red” has an evidentially modal conventional implicature, that would be one theory of the type I have been alluding to, a theory that locates a special feature in the concept expressed by words like “red” that is supposed to explain why they will be physically indefinable. So far as I know, no actual theorist has claimed that phenomenal concepts have this particular special feature: the example is intended only to illustrate the kind of thing that might be claimed for phenomenal concepts. For references to particular special-property accounts that have actually been made, about what have been called “recognitional concepts” or “quotational concepts” or the like, I refer the reader to Elpidorou.

If any one of these accounts is accepted — and the general line taken by such accounts has been widely influential, even if no one account of this type, based on a particular claim about the recognitional or quotational or whatever feature of phenomenal concepts, has a consensus in its favor — it would make a decisive point in favor of lacunism and against denialism, quite as much as would acceptance of the heuristic argument of the preceding section. In any case, I am going to dismiss denialism now, putting it in the same box with eliminativism, and restrict attention henceforth to lacunist versions of neuralism and psycho-functionalism.

But I have not yet mentioned the most striking feature of the phenomenal
concepts literature, sharply distinguishing it from my argumentation in the preceding section: most writers who propound phenomenal concepts arguments seem to regard them as triumphantly vindicating physicalism. The arguments are cited as if they disposed, not only of any problems raised by Jackson’s Mary, but also of any problems raised by Levine’s gap. Appeal to such arguments may be the most popular strategy among physicalists responding to Jackson or Levine. But can this phenomenal concepts strategy really dispose of the problem of the gap?

The strategy certainly does not dispose of any problem raised by the gap by disposing of the gap itself, filling it in or bridging it over. One can read all the literature on the topic listed by Elpidoorou and still have not even the beginning of an attempt at or an attempt at a beginning of a physical answer to the question, “Why does red look like this and not like that?” Quite the contrary, one will have been given more and more reasons to conclude that no physical answer ever will or could be given.

I suppose an extreme advocate of the principle that “ought implies can” might claim that the strategy thus establishes that it was a mistake ever even to wonder about the question. The thought would be that since the strategy establishes that the question cannot be answered, it was wrong to think that it ought to be answered. But such a line is hard to take seriously. How, then, is the phenomenal concepts strategy supposed to be of any help to physicalism in dealing with the problem of the explanatory gap?

Here Chalmers (2007), though himself a critic, has given an admirably clear and concise account of what the strategy is supposed by its proponents to accomplish. Chalmers writes that advocates of the strategy “agree that there is an explanatory gap, but hold that it stems from the way we think about consciousness. In particular, this view locates the gap in the relationship between our concepts of physical processes and our concepts of consciousness, rather than in the
relationship between physical processes and consciousness themselves.” For Chalmers such remarks are merely a preamble what he calls his “master argument,” an intricate discussion of the supposed beliefs of zombies, that is intended to undermine the phenomenal concepts strategy.

For a harder-core dualist, however, his prefatory remarks already strongly suggest that the strategy is simply misconceived. For according to Chalmers’ summary of what advocates of the strategy are up to, they are trying to give a physical explanation of “our epistemic situation.” If so, it must be remarked that it is all very well to say we have different kinds of concepts, and can consider one and the same thing under different ones of them, but one has not explained our epistemic situation until one explains how we can have concepts of the kind we have, concepts that there is something it is like or feels like to be considering. Anyone who thinks claims about phenomenal concepts provide a satisfactory physical explanation of our epistemic situation has misunderstood what that situation is (unless, as may be, in this discussion “our epistemic situation” is being revisionistically redefined to refer to something other than our epistemic situation as the phrase would be naturally understood; but it is the latter that needs explaining).

For our epistemic situation is that of conscious thinkers who engage in various episodes of conscious thinking, some of which are cases of doubting or erring, but some of which are cases of knowing or coming to know. Unlike thermostats and other electronic gadgets, we literally and consciously, and not just metaphorically, think and know. To provide an explanation of our epistemic situation, the physicalist would have fill or bridge the explanatory gap, not between sensation and affect and their neural correlates, but between conscious thought and its neural correlate. And to explain physically why we feel puzzled by the question “Why is seeing red like this?” one would first of all have to explain physically why
we feel anything at all and are not as unfeeling as thermostats. So far from doing
that, the literature on the phenomenal concepts strategy only gives reasons for
suspecting that it cannot be done.

But an analogy may be in order at this point. Imagine a museum with a
complex security system that among other things includes motion-detectors at the
periphery of its grounds, that when set off will, besides immediately doing things
like turn on flood lights, also transmit along a certain cable C an impulse to a
central control room. There something like PERIMETER BREACHED! will flash
in big, bright scarlet letters in the center of a screen, accompanied by the
trumpeting of an alarm.

We may suppose there is also an auxiliary system, routinely in operation in
the background, though mainly for use in trouble-shooting, which keeps a running
record of when this or that light is on or off, or this or that cable is transmitting or
failing to do so, and so on, rather as hospital monitors may keep track of the vital
signs of a patient. For there are external devices that can be used to detect whether
current is passing through a wire without those devices being connected to the wire
or on the receiving end of its current, just as we can imagine external devices
someday being developed that could be used to detect whether a patient’s C-fibers
are firing. The background system produces items like CABLE C TRANSMITTING
silently flashing in tiny, dim letters in a so-called crawl at the bottom of the screen,
as in TV news programs.

In this imagined set-up, a perimeter-breach signal being sent and cable C
transmitting an impulse are quite literally one and the very same event, but it gives
rise to two very different kinds of flashing on the monitor. Does this provide an
enlightening analogy to our having two contrasting concepts with no apriori
conceptual connection between them: a first-person concept of feeling pain and a
third-person concept of C-fibers firing? In one important respect it does not.
For the museum’s system, though a more sophisticated assemblage of gadgetry than a thermostat, has no concepts at all in the sense in which human beings in the control room watching the screen have concepts. Of what distinguishes human beings who think and feel from gadgets that merely flash, no physical account has been provided. While the museum example may provide a suggestive model of the neural correlate of our epistemic situation, it provides no model at all of that situation itself.

There is more to be said about the widely influential phenomenal concepts strategy, but for the moment at least perhaps enough has been said for purposes, not of settling its status, but of preliminary clarification of the scope of the physical. It is time to turn to the long-postponed discussion of modality.
5 Possibility and Reference

5.1 A Menu of Modal Flavors

In linguistics, **modality** is the distinction between what is affirmed as actual or real and what is not, insofar as it is grammaticalized: represented in language by either modifications of the main verb of a clause, called inflections of “grammatical mood,” or by use in addition to the main verb of certain other verbs, called “modal auxiliaries.” In English, distinctions of indicative vs subjunctive mood survive only vestigially, so it is mainly modal auxiliary verbs that are involved. The list of them overlaps with that of temporal auxiliary verbs, and consists of the four pairs “will, would, shall, should, may, might, can, could,” plus a few others. Four main flavors of modality in world languages are recognized by linguists, as surveyed by Frank Palmer (1986): evidential, considered in the preceding chapter and requiring no further consideration here, plus the trio of **deontic** and **epistemic** and **dynamic**. The terminology “deontic modality” and “epistemic modality” is borrowed from the literature of modal logic, as in the work of G. H. von Wright (1951). In English the same auxiliary verbs “may” and “must” are used with all three members of the trio, but in other languages they are expressed differently, and there can be no question but that there is a conceptual distinction to be drawn.

Deontic modality is illustrated by the contrast between “She may go,” and “He must stay,” meaning “She is permitted to go,” and “He is obliged to say.” This flavor will not be at issue for us here. Epistemic modality is illustrated by the contrast between “She must have seen it,” and “He may have seen it,” meaning “Given what is known, she must have seen it,” and “For all that is known, he may have seen it.” As knowledge varies from person to person and time to time, so does epistemic modality. The epistemic possibility of an afterlife — our not knowing for sure that there isn’t one — is the starting point for Blaise Pascal in his famous
wager. It is also as far as Descartes gets in his Second Meditation. But it is far short of Descartes’ ultimate goal, and again this flavor is not what will be at issue for us. As for dynamic modality, it pertains to what I or you or he or she or we or they can do or could have done, meaning “am/is/are able to do” or “was/were able to do.” Here, as in much of philosophy, what will be of most interest is an impersonal variant: “it could have happened that,” or “it could have been the case that,” where the “it” is pleonastic: the same it that does the raining when it rains.

Kripke, in the most famous single philosophical work on modality, Naming and Necessity (Kripke 1972, henceforth N&N), uses “possibility” without distinguishing adjective — possibility simpliciter, tout court, sans phrase — for this kind of possibility. For him the possible without distinguishing adjective is how things might have been if only, or what (is or isn’t but) could have been, and the correlative notion of the necessary without distinguishing adjective is how things would have been no matter what, or what (is and) couldn’t have failed to be. Kripke’s omission of any distinguishing adjective is at times inconvenient when so many others are using the modal vocabulary in so many other senses. Inspired by an off-hand remark of Kripke’s to the effect that what concerns him is a “metaphysical” rather than an epistemological notion — the scare-quotes are Kripke’s and are accompanied by an expression of a hope that the term will not be considered pejorative — in the literature this flavor of modality is often called metaphysical. That is the label that will be used here when a distinguishing label is needed; but usually I will leave it to be tacitly understood: metaphysical modality will be for us the default variety of modality.

In the literature of modal logic, deontic and epistemic, pertaining to duty and knowledge, are contrasted with alethic, the vanilla flavor, so to speak. The term etymologically means “pertaining to truth,” so alethic necessity is especially solid or firm truth. But the notion is ambiguous. For as W. V. O. Quine, a notorious
archenemy of modal thinking, whose status he denigrated especially in life-long debate with Rudolf Carnap, writes (Quine 1960: 59), “Philosophical tradition hints of three nested categories of firm truth: the analytic, the apriori, and the necessary,” by which last he means what is now called the *metaphysically necessary*. He adds, “Whether the first exhausts the second, and the second the third, are traditional matters of disagreement.”

It is a commonplace that the three notions are of conceptually different character: **necessity** as opposed to **contingency**, in the sense intended, has as noted been called “metaphysical,” pertaining to being; **aprioriness** as opposed to **aposterioriness** may be called “epistemological,” pertaining to knowledge; **analyticity** as opposed to **syntheticity** may be called “semantical,” pertaining to meaning. (Kripke remarks on these points early in N&N.) Note that while the epistemically possible was characterized earlier as what may be for all that is known, the apriori possible or epistemologically possible may be characterized as what may be for all that is knowable independently of sense-experience, and the analytically possible or semantically possible may be characterized as what may be for all that is knowable just from the meanings of words.

As to matters of traditional disagreement, as one moves from Leibniz to Kant to Frege and on to Carnap and P. F. Strawson and other contemporaries of Quine, there is a tendency for the necessary to be conflated with the apriori, and then the apriori identified with the analytic, and then the analytic explained in terms of semantic rules or linguistic conventions or the like. Quine himself famously attempted to persuade his colleagues do away even with this last, leaving none of the traditional categories standing. Quine did not persuade the majority, and on the contrary, one distinction formerly collapsed has since Quine’s heyday come to be restored to respectability, largely by examples of so-called aposteriori necessities in N&N. By contrast, there has been less tendency to revive the notion
of the synthetic apriori in a way that would be relevant to present concerns, and no such distinction will be important for us here. When apriori and analytic necessity are not distinguished, “conceptual necessity” (or “conceptual truth”) may be used as a neutral term, as if we weren’t dealing with enough vocabulary already.

This leaves us with two degrees of firm truth, the [metaphysically] necessary on the one hand, and the conceptually necessary on the other. A notion of firm truth brings with it a correlative soft notion of not being firmly untrue, and from some points of view it is more illuminating to start with the soft rather than the firm notions. In fact, we will have to take at least brief note of no fewer than four soft categories: the **suppositionally** (as I will call it) possible, the conceptually possible, the possible in the default or metaphysical sense, and the **nomologically** possible. The order in which I have listed them is from weakest to strongest.

The suppositionally possible is that which, unlike a round square or married bachelor, we can at least begin to consider without encountering anything obviously analytically or apriori false. The conceptually possible or **conceivable** is that whose falsehood is not analytic or apriori, not even unobviously so. The metaphysically possible is, as already indicated earlier, that which could have been the case. The nomologically possible is simply that which is metaphysically possible without any violations of any laws of nature that actually hold. There are candidate examples to distinguish each item on the list from the next after it.

Candidate examples of metaphysically possible nomological impossibilities are supplied in abundance by science fiction, and include gravity shields, faster-than-light travel, and more. According to dualist opponents of the supervenience thesis, zombies provide one more example: their existence is a metaphysical possibility, though one that violates natural laws of psychophysical correlation. The dualists with whom we will be concerned — epiphenomenalists — concede that the mental is nomologically supervenient on the physical, while denying that it
is genuinely or metaphysically supervenient.

Candidate examples of conceptually possible metaphysical impossibilities are supplied by N&N. In the literature, it is more common to speak of “aposteriori necessities” than of “conceptually possible metaphysical impossibilities,” but the notions are closely related: If \( p \) is necessary but aposteriori, not-\( p \) is impossible but conceivable, while if \( q \) is impossible but conceivable, not-\( q \) is necessary but not knowable apriori, and so either knowable but only aposteriori, or else unknowable.

We will be examining N&N a bit closer shortly, but for the moment we can just cite the most famous putative aposteriori necessity put forward there: the identity of Hesperus and Phosphorus, discovered in remote antiquity: Homer did not know of it, but contemporaneous Babylonian astronomers probably already did. The identity of Comet Hale and Comet Bopp, discovered in our own times, would be a parallel example. Kripke argues that, though clearly empirical discoveries, such astronomical identifications are [metaphysically] necessary.

As I have already said earlier, this book is not the place for any extended excursion into philosophy of mathematics. But mathematics does suggest a case of interest. It is often held, on the one hand, that mathematical statements are necessary if and only if they are true, and on the other hand, that they are apriori if and only if they are provable. Work of Kurt Gödel, Alfred Tarski, and others on the relationship of truth to proof in mathematics has suggested to many that there may be truths that are not provable — not merely unprovable from or relative to this or that accepted system of axioms, but in some sense absolutely unprovable. Such an unprovable mathematical truth would be a necessary truth not knowable apriori. It might simply be unknowable, or it might instead be well supported by computational verification of special cases and various kinds of heuristic argumentation, and so arguably knowable aposteriori. It is even sometimes suggested that Goldbach’s conjecture (to the effect that any even number greater
than four is a sum of two primes) may be an example of an aposteriori necessity of this kind. For illustrative purposes, we may pretend it is. This is not one of Kripke’s primary examples, but he briefly mentions it in passing in addenda to N&N as a case of a kind about which there had been some amount of discussion even prior to his own intervention.

5.2 Conceivability

While it may be reasonable in any given case of the superficially supposable to take there to be a defeasible presumption in favor of its conceivability, there are at least two kinds of cases where something turns out to be not coherently conceivable though it was superficially supposable. Both kinds call for discussion. One kind of case is easily described, but of relevance to our present concerns only for providing a contrast to the other case, which is more directly pertinent but less readily characterized.

The first kind of case is that of what is mathematically disprovable, but only by an unobvious disproof. It is superficially supposable, for instance, that there is a counterexample to Fermat’s famous conjecture. The mathematical gadfly Doron Zeilberger announced the discovery of one, without exhibiting it, in a blog post, and I am told that this was actually believed by some readers who did not note that the date of the post was April First. To be sure, it is provable that there are no such counterexamples, but the proof, due mainly to Andrew Wiles (who was awarded the Wolfskehl prize for it, alluded to in the opening paragraph of chapter 1) is long and difficult, and it was only after four centuries that “Fermat’s theorem” became a genuine theorem and not mere conjecture.

We can imagine a play or movie, on the order of David Auburn’s Proof (2000) about the discovery of a counterexample. We would be shown the consternation and turmoil among mathematicians that its supposed discovery
produces. We might be shown blackboards full of parts of the relevant calculations, rather as with the well-known Gary Larson Far Side cartoon of mathematical material on a blackboard with “then a miracle occurs” in the middle, and a figure pointing at it saying, “I think you should be more explicit here.” What we cannot picture is a whole explicit set of relevant calculations presenting a counterexample, since to picture that would be to produce a counterexample, and there are none.

The second kind of case (related to the so-called paradox of analysis in metaphilosophy) requires some background. To start far from contentious issues, all native English speakers can effortlessly produce the plurals of English nouns in the regular case, excluding talk of geese or oxen or the like. But most cannot state the rule for when one adds an s-sound, as with “gates,” or a z-sound, as with “gales,” or a syllable consisting of a neutral vowel followed by a z-sound, as with “gazes.” Most do not even have in their vocabularies the IPA notation or the technical terms in which phonologists would formulate the rule: “voiced, sibilant, schwa,” and so forth.

I have said that there is much more to philosophy of mind than the mind-body problem, and certainly there is more to psychology than issues about consciousness. One instance is that linguists often hold that rules such as one for pluralization are in some sense psychologically real; but though native speakers in some sense follow them, they do not consciously follow them the way someone learning English as a foreign language might. Nor is there any psychoanalysis to bring unconscious phonological rules to consciousness.

All speakers are aware of consciously, and the only so-called linguistic intuitions they can supply as data to the linguist who consults them as informants, are case-by-case judgments about what is the plural of this word, or that word, or some other word. The linguist seeking to arrive at a rule can only proceed by
looking at a body of such judgments, forming a hypothesis that seems to fit them, testing this original hypothesis against further such judgments, revising the hypothesis if discrepancies are found, and so on until the process stabilizes. In the case where the linguist is concerned only with his or her own idiolect, with him- or herself as sole informant, this is an armchair process; in other cases, it is a form of social-science research.

The situation seems entirely similar with grammatical and semantical as contrasted with phonological rules. Lexicographers seeking definitions and philosophers seeking analyses start with candidates, check these against intuitions about particular examples, revise, and repeat. The need for such a dialectic is why lexicography is as much art as science, and why expertise in the use of words does not inevitably entail facility in producing definitions. Though Shakespeare was an immeasurably greater writer than Samuel Johnson, we cannot be confident that if he had set about to produce a dictionary he would have produced a better one. The fact that semantic rules can only be got at indirectly is why in philosophy correct analyses may not be immediately recognized as right, and why erroneous analyses may not be immediately recognized as wrong. A famous case is that of the analysis of knowledge as justified true belief.

Another arguable case is presented by analyses according to which pain consists of a sensation plus an affect. For there are drugs that can break the sensation-affect link. I myself have only once experienced the effect of such a drug, and found myself quite surprised by it. I described my state as “feeling pain but not being bothered by it,” and I am told that others who have had similar surprising experiences have described them in similar ways. The ordinary childhood training in the use of the word “pain” seems to leave the speaker disposed to describe the unusual experience in this way. If so, then it seems that the ordinary meaning of the word “pain,” as internalized after the ordinary childhood
training in the use of the word, is such that pain can properly be said to be felt when only the sensation, and not the affect that usually accompanies it, is present. If so, then it seems that “pain” denotes the sensation, rather than the affect or the complex of sensation plus affect.

There is a point about meaning here that needs emphasis, namely, that when checking an example, when considering whether such-and-such an expression would be appropriate in such-and-such circumstances, one normally does not need to have been in or need to get oneself into those circumstances. Usually it is enough just to imagine them. But it is a peculiarity of sensory experiences that those who have never had them generally cannot imagine them, either. When philosophers who have never experienced a situation where the sensation is present but not the affect try to guess what would be the correct use of “pain” in such a situation, they may go wrong. That is to say, they may arrive at judgments differing from those displayed by ordinary speakers who are actually experiencing such a situation. They may guess that without the affect the sensation cannot properly be called “pain,” and so offer definitions or analyses according to which pain is inherently unpleasant; and indeed such definitions and analyses are quite common. Arguably such analyses are simply wrong.

Against all this it may be urged that the expert authorities, scientific groups devoted to the study of pain, have offered official definitions according to which pain is an unpleasant sensory experience, building the affect in. To this argument from authority it must be replied that, while specialist groups may have very good reasons for framing their definitions as they do for their particular purposes, it does not follow that those technical definitions are in agreement with the ordinary meaning of the word. Across the sciences, technical uses of ordinary terms rarely preserve fully their original meanings. Nor is it obvious that where technical and ordinary usage diverge, philosophers ought to follow the specialist rather than the
lay practice.

A clear counterexample is provided by the case of “abortion,” where professional medical terminology radically diverges from popular usage. The latter distinguishes abortion from miscarriage, while the former calls both “abortion” and distinguishes them as induced and spontaneous. Philosophers generally and rightly follow the populace rather than the medicos, and though there is a vast philosophical literature indexed under the term “abortion,” what is at issue in this literature is “abortion” in the popular sense, which in medical terms would be called “induced abortion.” There is almost no philosophical literature on abortion in the medical sense, which in popular terms would be called disjunctively “abortion-or-miscarriage,” since induced abortion and miscarriage seem to have next to nothing of philosophical, as opposed to medical, interest in common.

If it is not easy to imagine pain without suffering, it was perhaps even more difficult to imagine justified truth belief without knowledge, and Edmund Gettier’s paper (1963) presenting an example is justly celebrated. According to analytic functionalism, we tend to fall into the opposite sort of problem with pain: not being taken in by an analysis which is wrong, but failing to accept an analysis that is right. We do not recognize that to be in pain simply is for there to be a state realizing the role of being caused by tissue damage and causing avoidance behavior, and our being in it. We seem to be able to imagine zombies who are in states realizing the relevant role but feeling nothing at all, neither pain nor anything else; but according to the analytic functionalist, this is a mistake.

It is only the decision, early on in this study, to dismiss analytic functionalism that allows us to neglect the gap between suppositional and conceptual possibility henceforth, except in incidental asides. The physicalists with whom we will be concerned — lacunists, usually of the neuralist variety — concede that the mental is not “conceptually supervenient” on the physical, while
insisting that it *is* genuinely or metaphysically supervenient.

So far as the debate over supervenience is concerned, we are left with following question: on which side of the line separating the metaphysically possible from the metaphysically impossible does the zombie universe fall? The dualists with whom we are concerned grant their opponents’ claim that zombies are nomologically impossible, but challenge them to show that they are metaphysically impossible. Sci Fi provides epiphenomenalists with many examples they might cite as cases of nomologically impossible metaphysical possibilities. They challenge their opponents to explain how zombies are supposed to differ from these. The physicalists with whom we are concerned grant their opponents’ claim that zombies are conceptually possible, but challenge them to show that they are metaphysically possible. N&N provides neuralists with several examples they might cite as cases of conceptual possibilities that are metaphysically impossible. They challenge their opponents to explain how zombies are supposed to differ from these.

There is a certain symmetry here, but as I have said earlier, in the opposition between physicalism and dualism, the former is so much more the favorite with the philosophical public than the latter that “it is often treated as the default view, as if the burden of proof were obviously on the other side.” This makes for an asymmetrical literature, in which the issue of how to get from coherent conception to metaphysical possibility is much more often addressed than the issue of how to get from natural law to metaphysical necessity. The former rather than the latter question will accordingly be the almost exclusive focus of what follows here.

But let me begin by reiterating what it is the physicalist does and does not concede. The main concession is that the correlation between being in pain and having C-fibers that are firing is an empirical, aposteriori, synthetic discovery. It is not a conceptual, apriori, analytic truth that everything mental has a physical
correlate at all, and that the soul of the sleeper is not capable of slipping away from
the body and engaging in independent travels to a spirit realm beyond the material.
Nor is it a conceptual, apriori, analytic truth that if the mental has a physical
 correlate, that correlate must be in the body, as opposed to the body’s being a
remote terminal, while thinking goes on in a cloud. Nor is it a conceptual, apriori,
analytic truth that if the mental has a bodily correlate, it is in the brain rather than
some other organ, the heart or liver or spleen or kidneys.

But these comparatively uncontroversial observations only establish the
conceivability of the failure of mind-body correlation, and more in the minds-
without-bodies direction than the opposite. I am taking for granted that by now
more has been conceded, namely, that it is conceivable that everything neural or
encephalic, and more generally all bodily states and processes, and more generally
still the whole material and physical world, could have been just as it is, yet mental
life absent. Yet once N&N has convincingly argued that there can be things
conceivable yet impossible, the concession that the zombie universe is coherently
conceivable is no longer enough for the purposes of the dualist, who wants to claim
it is genuinely possible; and this is what the physicalist will not concede.

5.3 An Un-Kripkean Approach

Two substantial bodies of work by authors writing from a point of view at
least partially sympathetic to some kind or other of dualistical view are available,
both relevant to how the dualist might respond to the challenge of the gap between
conceivability and possibility. On the one hand, there is the work of Kripke
himself, especially towards the end of the last of the three lectures in N&N, and in
the supplementary material added in the form of footnotes and an afterword when
the transcript of the audio tape of the lectures was published. The material in the
lectures generated an enormous amount of discussion; the supplementary material,
a good deal, if not as much. The Saul Kripke Center at the City University of New York Graduate Center has many further audiotapes in its files, awaiting digitization and transcription and editing for publication, but whatever additional relevant material may be in them has as yet played no role in debate over the mind-body problem. In the absence of substantial new published material on that problem from Kripke, discussion of his work has somewhat subsided, and the work itself has tended to be partially eclipsed for many by newer work of others.

On the other hand, there is the work of the enormously prolific Chalmers, the foremost among these others, going back to his influential book on the conscious mind, a quarter-century (1996) after Kripke’s famous Princeton lectures that became N&N. Chalmers’ personal webpage has been a hub for information on the topic, with links not only to his own work published and forthcoming, but to that of many of his most important critics, along with his replies. Jackson’s work must also be mentioned. For his position as of around the time of Chalmers’ book was dualist — there has since been a falling away (in Jackson, 2004) — and similar to that of Chalmers: not identical, but similar enough that in short discussions in the literature often little distinction is made between the two. How different the Kripkean approach is from the Chalmers-Jackson approach is recognized by some, notably David Papineau (2008), but seemingly ignored by many others.

One point of agreement between Kripke and Chalmers seems to be — and I infer from personal communications that Jackson agrees as well, even now that he has abandoned dualism — that however convincingly Kripke may have argued that there are examples of conceivable impossibilities, his arguments do not provide any general license to posit new instances of the phenomenon at will, wherever it would suit a philosopher’s convenience or confirm a philosopher’s prejudices. To say this is as much as to say that the burden of proof does not, after all, lie entirely on the dualist side. And the efforts of all three of Jackson, Chalmers, and Kripke
can be viewed, and are perhaps best viewed, not as efforts to make a knock-down case for dualism, but rather as substantial attempts to shift the burden of proof to the physicalist side. Looked at from this point of view the approaches share a common structure.

Each approach first identifies some key feature, claims that any example $Q$ of an alleged aposteriori necessity put forward by Kripke in N&N has the feature, and then claims that some example $P$ of an alleged aposteriori necessity put forward by neuralist proponents of supervenience lacks the feature. We may as well take $P$ to be the old stand-by:

(P) Being in pain is correlated with having C-fibers that are firing.

As for $Q$, the only examples given so far have been Goldbach’s conjecture, which plays a very marginal role in Kripke’s discussion, and identity statements involving heavenly bodies, which are not really typical, either, but could find a place as a representative of the first of a half-dozen main types of examples he considers, as illustrated by these:

Hesperus and Phosphorus are the same.
Gorse plants and furze plants are the same.
Nero was descended from Augustus.
Whales are descended from therapsids.
The Colossus of Rhodes was composed of copper, among other ingredients.
Sugars are composed of carbon, among other ingredients.

Except for the first, these are “Kripkean” in the sense of being variations on examples in N&N rather than “Kripke’s” in the sense of being taken directly from that work. The examples come in pairs consisting of a case pertaining to an
individual item and one pertaining to a kind of item. Three special relationships are involved: in the first pair, identity or being the same as; in the second pair, pedigree or being descended from; in the third pair, constitution or being in part composed of.

For present purposes it will be convenient to consider yet another example (which could be construed as a variant type of constitution example):

(Q) Tin is composed of atoms having fifty protons each.

An example like this would in the literature usually be quoted in a form like “Tin has atomic number 50,” but it may be well to spell out what atomic numbers are now understood to denote, namely, the number of protons per atom of an element. The dualist’s goal will be to identify a key feature, show that P lacks it, and show that Q has it, moreover showing this last in a way that would apply also to any other Kripkean example. If all this can be achieved, the dualist will have put the ball back in the physicalist’s court.

One obvious difference between Q and P is that Q does, while P does not, involve overtly phenomenal or mentalistic vocabulary. But it would be clearly question-begging to cite this difference, so the key feature to be identified will have to be something else.

Chalmers and Jackson rely on a considerable amount of what has proved to be highly contentious apparatus (called “two-dimensionalism”), presenting the expositor with a dilemma. On the one hand, an attempt to describe the Chalmersian and Jacksonian arguments that ignores their theoretical apparatus cannot be wholly faithful to the viewpoint from which the arguments are given, to put it mildly. On the other hand, any attempt to deal with the contentious issue of the status of that apparatus in a book no longer than this one will almost inevitably be inadequate, to
use no stronger term. The dilemma arises regardless of the expositor’s personal opinion about the apparatus, though for me, sharing as I do some of the kinds of doubts about the apparatus expressed at book length by Scott Soames (2004), the first horn is the less repellant, and I will impale myself on it: I will attempt an apparatus-free sketch of an un-Kripkean dualistical argument, inspired by Chalmers and Jackson; but I acknowledge in advance that my criticisms of un-Kripkeanism may not directly apply to the overall position of either writer from whom inspiration has been taken.

Every species of necessity carries with it not only a correlative species of possibility, but also a species of implication or consequence In particular, \( B \) is a conceptual implication or consequence of \( A \) if the conditional statement “if \( A \) then \( B \)” is conceptually necessary. Let now \( \Omega \) be the conjunction of all truths expressible in physical terms. Don’t worry about the infinity of the relevant set of truths: replace “all” by “enough,” if you will. The pro-lacunist, anti-denialist arguments considered earlier are supposed to show that \( Q \) is not a conceptual consequence of \( \Omega \). As for \( P \), it is trivially a conceptual consequence of \( \Omega \), since it is a conjunct of \( \Omega \), what with its being true and all its vocabulary being physical. But the fact that \( P \) is while \( Q \) is not a conceptual consequence of \( \Omega \) amounts to little more than the fact that \( P \) does not while \( Q \) does involve phenomenal vocabulary, a fact whose cogency as an objection has already been dismissed.

The un-Kripkean strategy I want to consider involves distinguishing a significantly restricted subset of the total physical vocabulary, to be called basic, and considering, instead of \( \Omega \), the conjunction of all truths in physical vocabulary, only \( \Omega_0 \), the conjunction of all truths in basic physical vocabulary. With \( Q \) not being a conceptual implication of \( \Omega \) it will a fortiori not be a conceptual implication of \( \Omega_0 \), but the claim will be that \( P \) still is a conceptual consequence of the weaker \( \Omega_0 \) and not just the stronger \( \Omega \). In other words, it is being a conceptual
consequence of basic physical truths that is the key feature on this approach.

In trying to evaluate this approach, it may be well to go slowly, and first review the grounds for the generally accepted conclusion that \( P \) is not a conceptual truth, or part of the very meaning of “tin.” Here one should not be misled by the fact that a present-day dictionary might include the information \( P \) in its entry for the word “tin.” For by so doing, the dictionary is stepping out of the role of a lexicon or glossary and into the role of an encyclopedia or handbook, providing nonlinguistic information. For ignorance of \( P \) surely is not linguistic ignorance.

Dmitri Mendeleyev published his version of the period table, from which his fame and the assignment of atomic numbers both ultimately derive, only in 1869; and the present atomic numbering of elements, taking into account such noble gases as neon, dates only from the 1890s; and the interpretation of atomic numbers in terms of subatomic structure and numbers of protons came later still. Charles Dickens died in 1870, after a career of several decades producing a corpus of literature that, according to on-line concordances, uses the word “tin” five dozen times. If ignorance of \( P \) were ignorance of the meaning of “tin,” we would have to conclude either that Dickens didn’t know what the word meant, which is absurd, or that the meaning of the word has shifted between his time and ours. But dictionaries themselves implicitly testify against this latter assumption, since it is their custom, when a word changes meaning over time, to list the successive meanings by separate numbers under the word, and this is not done in the case of “tin.” No other characterization of tin based on microscopic features can be a conceptual truth, either, for similar reasons.

So much for why \( P \) is not a conceptual truth, but what about its being a conceptual consequence of \( \Omega_0 \)? To evaluate claims about \( that \), we must look at the vocabulary in terms of which \( \Omega_0 \) is formulated. It is reasonable to assume that basic vocabulary would include both notions pertaining to physical microstructures and
notions pertaining to the macroscopic spatiotemporal distribution of such structures, relative to our location here and now, so that in particular the following would, if true, be a conjunct and hence a conceptual implication of $\Omega_0$:

(1) The second ingredient in the alloy most often produced on the surface of this planet in the period from 5000 to 3000 years before the present is composed of atoms having fifty protons each.

And I take it (1) is true, since tin, which is composed of atoms having fifty protons each, is the second ingredient after copper in bronze, and the Bronze Age began about 3000 BCE with the Iron Age following about 1000 BCE. For $P$ to be a conceptual consequence of $\Omega_0$, it would be enough that the following should be one also:

(2) The second ingredient in the alloy most often produced on the surface of this planet in the period from 5000 to 3000 years before the present is tin.

But it is hard to see how such a characterization, with its obvious dependence on archeological excavation and metallurgical analysis, could be claimed to be part of the very meaning of “tin.”

Now any other characterization in terms of basic physical vocabulary would do, if only some such characterization could be claimed to be part of the very meaning of “tin.” We would have in place of (1) and (2) this pair, wherein “tinny stuff” is a placeholder for some characterization in basic physical vocabulary:

(3) The tinny stuff in the vicinity of here and now is composed of atoms having fifty protons each.
(4) The tinny stuff in the vicinity of here and now is tin.
Just at this point, however, we come to the key disagreement between Kripke on the one hand and un-Kripkeans on the other. In order for (4) to be a conceptual truth, the un-Kripkean needs there to be, at least for each speaker on each occasion of use, some uniquely identifying description of tin, represented here by the placeholder phrase “tinny stuff,” associated conceptually with the word “tin” as something very like a meaning. Kripke, by contrast, denies there is any such thing.

The disagreement here extends to two large classes of expressions: first, such phrases as “Murray Gell-Mann” or “Steven Weinberg,” generally called “proper nouns” or proper names; second, such words as “tin” and “zinc,” belonging to a subclass of what are generally called “common nouns,” which subclass are sometimes called common names or “natural-kind terms.” The relation that the names “Murray Gell-Mann” and “Steven Weinberg” and “zinc” and “tin” bear to Murray Gell-Mann and Steven Weinberg and zinc and tin is reference, a.k.a. “denotation” or “designation.”

Since the nineteenth century, this notion has been contrasted with what is called sense, a.k.a. “connotation” or “meaning,” where it is understood that expressions may have the same reference while differing in sense, as “the most famous student of Socrates” and “the most famous teacher of Aristotle” have very different connotations or meanings, but the same denotation or designation, Plato. The terms “sense” and “reference” come from English translations of Frege, and Kripke’s main target in his famous lectures is a view, descriptivism, associated with the name of Frege, according to which names, both proper and common, have uniquely identifying descriptive senses. And so we are led deeply into contentious issues of philosophy of language.

Some Kripke-inspired critics of the Chalmers-Jackson apparatus, such as
Alex Byrne and James Pryor (2004), in effect take its advocates to task for failing to appreciate the significance of Kripke’s arguments — especially those that turn more on epistemic than modal considerations — against descriptivism, or for failing to appreciate the similarity of the central notion of their apparatus (called “primary intension”) to Frege’s notion of sense. Defenders of the apparatus respond, attackers rebut, defenders rejoin, and a large literature results, none of which will I directly discuss, though my discussion of Kripke’s antidescriptivism to follow will be influenced by bits and pieces of it.

5.4 Modality and Reference

Kripke presents his antidescriptivism first for proper names, then for common names, and then discusses applications. I will follow suit in my summary here. Kripke has two kinds of arguments which may be called the modal arguments and epistemic arguments, the latter including the argument from ignorance and the argument from error. In connection with the epistemic arguments Kripke acknowledges overlap with independent work of Keith Donnellan — and of precursors back to the late 1950s, though these only saw a need to amend descriptivism, not to abandon it. In connection with the extension from proper to common names Kripke acknowledges overlap with independent work of Putnam.

Both the modal and epistemic arguments are in the first instance directed against the thought that a proper name C might be associated with a uniquely identifying description of the form “the fullfiller of condition 𝜓,” say “Aristotle” with “the teacher of Alexander,” or “Murray Gell-Mann” with “the particle physicist who developed the theory of strangeness.” The modal argument is that if the name and the description were alike in meaning, then the following would also be alike in meaning:
(1) It could have happened that Aristotle did not teach Alexander.
(2) It could have happened that the teacher of Alexander did not teach Alexander.

But whereas (1) is unambiguously true, since Aristotle might have taken up his father’s profession of medicine and never gone into teaching at all, by contrast (2) is ambiguous between a true reading and an absurd one:

(3) Someone taught Alexander but it could have happened that he did not teach Alexander.
(4) It could have happened that someone taught Alexander but did not teach Alexander.

In speaking of an alternate course history might have taken, say one in which some rival of Aristotle got the job of teaching Alexander, the description “the teacher of Alexander” might be read as referring to the person who in the actual course of history did teach Alexander, which is to say, Aristotle, giving reading (3), or the person who in the alternate course of history would have taught Alexander, which is to say, the rival, giving reading (4). But the name “Aristotle” refers to Aristotle whether speaking of the actual or the alternate course of history: as is said, the reference of the name is rigid, while the reference of the description is flexible. And that should be enough to show they are not synonymous.

The epistemic argument is that one can be able to use the name “Gell-Mann” to refer to Gell-Mann even if one can provide no true uniquely identifying description of him. All X and Y, respectively, may able to say may be this:

(5) Gell-Mann is a famous physicist.
(6) Gell-Mann is the physicist who unified the electromagnetic and weak forces.

But (5) is equally true of Gell-Mann and Weinberg, while (6) is false of Gell-Mann but true of Weinberg. Yet X and Y, for all their ignorance and error, are both referring to Gell-Mann, one saying something true but not very informative about him, the other saying something false about him.

So even if we allow that the meaning of a name may vary from person to person and time to time, no uniquely identifying descriptive meaning need be associated with the name by a given person at a given time. Exactly similar considerations apply to common names: X may only be able to say that tin is a grayish metal or semimetal often used for plating or in alloys, which does not distinguish it from zinc. Y may only be able to say that tin is the second ingredient after copper in the alloy brass, which is false of tin but true of zinc.

If an associated uniquely identifying descriptive sense does not determine what person or substance or other item a proper or common name refers to, what does? Kripke offers a picture of a chain of communication, on which an item is picked out — say by pointing at a person, or at a sample of a substance, or whatever — and a name assigned it by what he calls an “initial baptist,” and then the name is passed from speaker to speaker, each adopting it with the intention of using it to refer to what the speaker from whom the name was adopted was using it to refer to. No uniquely true noncircularly-identifying information need be passed on along with the name.

Against the modal argument it is often suggested that we could take “Aristotle” to mean, not “the teacher of Alexander” but “the actual teacher of Alexander” or “the person who actually did teach Alexander.” By inserting “actually” we rigidify the description, as is said. This knocks out one reading of “It could have happened that Aristotle did not teach Alexander” leaving only the
reading on which “Aristotle” rigidly denotes the same person in the hypothetical contrary-to-fact situation as in the actual situation.

Against the epistemic argument it is often suggested that there is at least one thing X and Y could say that would be uniquely true of Gell-Mann.

(7) Gell-Mann is the bearer of the name “Gell-Mann.”

One really should add here “as used in the present discussion,” or something of the sort, since presumably Gell-Mann has relatives who share the family name; but let the phrase “as used in the present discussion” be tacitly understood, and grant that (7) is uniquely true of Gell-Man. Still everyone’s associating with the name “Gell-Mann” the description (7) cannot suffice to identify who Gell-Mann is, any more than the inhabitants of an island can all support themselves by taking in each other’s washing. The adherent of the metalinguistic theory, as one calls the view that a name \( N \) means “that which bears the name \( N \),” will need some account of what determines the reference of names, if not a descriptive sense associated with them. But there is no obvious reason why such an adherent cannot simply appropriate Kripke’s chain-of-communication picture.

A little thought shows that rigidifying by inserting “actually” does not help with the epistemic problem: rigidifying an identifying description may avoid modal problems, but what if one is so far sunk in ignorance and error as to lack any true uniquely identifying description to rigidify? A little thought also shows that going metalinguistic does not help with the modal problem, since a different person might have borne a given name than the one who actually does bear it. If we put on both patches at once, however, we are left with one surviving candidate as the potential descriptive meaning of “Gell-Mann” or “tin,” thus:
(8) the individual who actually bears the proper name “Gell-Mann”
(9) the substance that actually bears the common name “tin”

There are still problems for the descriptivist, though I will not press them, but rather will conclude my excursion into general philosophy of language here.

The question left hanging earlier was that of whether tin admits a definition that, when added to a complete description $\Omega_0$ of the world in basic physical terms, would imply that tin is the element whose atoms have fifty protons each. Our recent discussion leaves only one candidate definition standing, that tin is the kind of stuff that actually bears the common name “tin,” or more simply:

(10) What “tin” actually refers to is tin.

(What I was earlier calling tinny stuff now becomes stuff actually bearing the name “tin.”) What we now want to know is whether this is a conceptual implication of $\Omega_0$:

(11) What “tin” actually refers to is composed of atoms having fifty protons each.

We may suppose $\Omega_0$ to include or imply information about what kinds of sounds are emitted by what kinds of organisms where, and so on, including information about the utterance of “tin” here and now, on the earth of the twenty-first century. The question thus is whether when conceptual truths about reference are added to such information, (11) follows. That depends ultimately and crucially on whether some kind of physical analysis of the notion of “reference” can be produced. Kripke himself is adamant that his chain-of-communication picture is only a picture, and not a reductive analysis permitting “reference” to be defined in physicalistic terms. Could something more be made of the chain-of-
communication picture than Kripke himself does or thinks can be done? If so, there
may be some hope of sustaining a physical analysis of reference; if not, the
un-Kripkean line of dualistic argument seems to falter. That is perhaps as far as the
issue can be followed without getting deeper into contentious issues about
theoretical apparatus, Kripkean and Chalmersiano-Jacksonian.

So I will now set aside the un-Kripkean approach and turn back to Kripke
and our contrasting examples:

(P) Being in pain is correlated with having C-fibers that are firing.
(Q) Tin is composed of atoms having fifty protons each.

How might Kripke distinguish them? His discussion, which though addressed to
other examples would apply mutatis mutandis to ours, suggests that in the case of
Q there are explanations both of (i) why it may seem to be contingent and of (ii)
how we can know that it isn’t. And Kripke further suggests that these kinds of
explanations simply would not be available in a case like P. The availability of
such explanations is the key feature distinguishing Kripkean examples from
neuralist pseudo-examples on this approach. Kripke’s account of the false
appearance of contingency is given in the third lecture, and his account of how we
can have knowledge of necessity is given in the addenda to the lectures. The
former account goes into more detail than the latter, but both are significant.

To begin with false appearances, one to whom it appears that Q is contingent
might say, “When investigation was undertaken into what tin is composed of, it
might have turned out to be something other than atoms having fifty protons each:
investigation might have discovered, say, a non-atomic composition.” The
Kripkean answer must be, “No, investigation could not have discovered a non-
atomic composition, any more than the voyage of Columbus could have discovered
Atlantis. There was no Atlantis there to be discovered, nor any composition of tin out of anything but atoms having fifty protons each.” What is true is that when the investigations began, *for all that was known at the time*, it might have been heading towards the discovery of a non-atomic composition, since it was not known at the time that tin did not have one. All this is as much as to say that “it could have turned out that” is best understood as an idiom of *epistemic* rather than and as opposed to *metaphysical* modality.

Kripke, however, is not content just to say that because English uses very similar forms of words to express epistemic and metaphysical modalities, it is easy to confuse them. Rather, he attempts to explain the mistaken feeling that certain impossibilities are possible as being a result of confusing those impossibilities with certain genuine possibilities. He insists that one must be careful to distinguish what a word is used for by us when speaking of a counterfactual situation and what it would be used for by creatures speaking in a counterfactual situation, supposing creatures capable of speech would exist in that situation. He then points to confusion on just this point as responsible for what he takes to be the mistaken feeling that there is some counterfactual situation in which tin would have been revealed by investigation to have a non-atomic composition.

According to Kripke, what there really is, is a counterfactual situation or so-called possible world in which there are creatures in the same epistemic relation to what they call “tin” as we were in to what we call “tin,” which is to say, as we were in to tin, before certain investigations were undertaken, but in which subsequent investigation revealed that what they were calling “tin” has a non-atomic composition. On the Kripkean view, however, this would not be a world in which tin has a non-atomic composition, but rather a world in which “tin” is used to denote something other than tin.

The difference between $Q$ and $P$ here is that while there could have been
creatures in a situation epistemically indistinguishable from ours prior to certain discoveries, involving something they called “tin” though it was not tin, there could not have been creatures in a situation epistemically indistinguishable from ours prior to certain discoveries, involving something they called “pain” though it was not pain. There could, according to Kripke, be what may be called “fool’s tin” — analogous to “fool’s gold” — but no “fool’s pain.” That is because nothing could be epistemically indistinguishable from pain unless it felt like pain, and anything that feels like pain is pain, since pain is a feeling.

Nathan Salmon (1981) long ago emphasized that among Kripke's examples of aposteriori necessities in N&N, only the Hesperus = Phosphorus example and others pertaining to identity depend only on his views about reference and rigidity. His case for the pedigree and constitution examples involves the evocation of additional intuitions. If we look carefully at the implicit background in the Kripkean analysis as I have sketched it, we find two kinds of assumptions.

First, there is a rigidity assumption according to which a word like “tin” or “pain” in our mouths denotes the same thing when speaking of a counterfactual situation as it does when speaking of the actual situation. But second, there is a view about what it is for something in a counterfactual situation to be or to count as the same thing as tin or pain in the actual situation. This seems to be a view involving the classification of tin as a substance and pain as a feeling, linked to the assumption that a substance in another possible world counts as the same substance as a substance in our actual world just in case it has the same composition, while a feeling in another possible world counts as the same feeling as a feeling in our actual world just in case it feels the same. What we find in the background assumptions will it seems be much the same when we turn from false appearances of contingency to sources of knowledge of necessity.

It is fairly common for metaphysicians today, especially of the school who go
in for enthusiastic and exuberant speculation about what they call “essence” and “fundamentality” and “grounding,” to be rather unconcerned with questions about how we can know this or that metaphysical claim is true. One well-known leader in the field has even been quoted as coming out with the slogan “Epistemology last, if ever,” though as I have not been able to find that phrase in print, I will name no names. For all one could tell from the text of his three lectures, Kripke might have been of a similar way of thinking; but the addenda provided when the lectures were published as N&N show that the question “How do we know?” was never far from his mind, even if his partial answers are extremely cautious and heavily guarded, presenting not solutions but what he says are hints of what may be a solution.

In part (b) of the addenda he quotes the well-known line from Kant in the introduction to the second or B version of the Critique of Pure Reason, according to which “Experience teaches us that a thing is so and so, but not that it cannot be otherwise,” criticizing Kant’s conclusion that empirical observation is never required to establish a necessary truth, while granting his premise that it is never sufficient to do so. He makes the important observation that in all his basic examples of aposteriori necessities, though empirical observation is required to establish the truth of the example, “philosophical analysis” tells us in advance that if it is true then it is necessarily so.

(Parenthetically, this is not to say that every case of an aposteriori necessity will be like this. On the contrary, given a couple of independent basic Kripkean examples plus any unrelated ordinary, contingent, aposteriori truth, it is easy enough to concoct a truth-functional compound that is necessarily true but such that, for all that can be known apriori, it may be necessarily true, contingently true, contingently false, or necessarily false. I leave the production of an example as an exercise to the reader.)

Conceptual truths have been widely taken to be consequences of semantic
rules or linguistic conventions, which would seem to reduce knowledge of them to knowledge of the rules and conventions of our own speech, thus rendering it considerably less mysterious than if it had to be regarded as acquired through some occult faculty of intuition. One may wonder, however, what the rule or convention might be in the case of this or that example. Though Kripke may not state any such underlying rule or convention explicitly, it seems not hard to guess what it would have to be in the tin case, namely, something like this:

(1) If a substance has a certain composition, then any substance there is or could have been with a different composition does not count as the same substance.

By contrast, if we search for a key feature to distinguish the pain case from the tin case, the Kripkean claim would presumably be that there is nothing in the background comparable to (1) above in the way of a conceptual truth that would add an element of necessity to the empirically discovered correlation of mental with neural states, boosting aposteriori truth to aposteriori necessity. For it seems clear what would be needed, namely, something like this:

(2) If a feeling has a certain neural correlate, then any feeling there is or could have been with a different neural correlate does not count as the same feeling.

And the Kripkean contention would be nothing at all like (2) can be plausibly maintained, because what makes conscious states count as the same are such factors as who is the subject of them and when they occur and what it is like or feels like to be subject to them, not what neural goings-on are underway when they occur. One would have to be grossly mistaken about what makes for sameness of feelings to think that something like (2) above held as a semantical rule or linguistic convention.
But it is only too obvious now that if my pessimism about making non-Kripkean arguments work is at all appropriate, leaving us only with Kripke’s approach, and if my account of what is going on in the foreground and background of Kripke’s discussion is anywhere near accurate, then the dualist case is left depending on exactly the same kind of assumptions about classification and identity criteria that would be involved in directly arguing that the neuralist’s (or separately, but in just the same way, the functionalist’s) identity claims are conceptually confused. If direct debate of this kind threatens to reach an impasse, as I suggested earlier, then the detour through modal considerations has not after all taken us around it, but rather has brought us right back to it after forcing us to make a long circuit.

Moreover, taking the modal route may be not only futile but dangerous. For we have been exposed all along the way to the threat of becoming bogged down in obscurities about modal apparatus, a threat I have tried to ward off here mainly by playing down the role of any theory of so-called possible worlds that takes talk of them to be anything more than a picturesque manner of speaking, and not wading into the debate over apparatus between the un-Kripkeans Chalmers and Jackson on the one hand, and Kripke supporters like Soames or Byrne and Pryor on the other. But if we have not sunk in modal quicksand, still we are left with the question unsettled whether there is any way to decide the issues left outstanding that does not involve appeal to intuitions about identity criteria — and with the question how important it really is to decide such issues at all.
6 Reflection and Projection

6.1 Status Quæstionis

To repudiate denialism is to acknowledge that the physical account of the world is radically incomplete — and not just the account that is offered by physics in the narrow sense, but the account that is offered by physical science in the broadest sense, and not just the account that is offered by actual physical science today, but any account that might be offered by any potential physical science of the future. In particular, it is to concede that even “the best idea anyone ever had” (as Dennett calls Darwin’s idea of natural selection) is powerless to explain the most familiar fact of all, that we have mental as well as physical lives.

Considerations offered in the preceding chapter suggest that the incompleteness noted derives from the very nature of our concept of the physical as it has developed since the seventeenth-century ban on secondary qualities and/or from the very nature of our phenomenal concepts, making it virtually a conceptual truth that there can be no physical explanation of why we have feelings, nor any physical description of what it is like to be having the feelings we have, nor any physical explanation of why it is like that.

To repudiate interactionism is to acknowledge, however, that what physical science leaves out of its account of the world is of no importance for physical science itself: what is left out has no effects describable in physical terms, since by the principle of uniformity, to the extent that what is physically describable is subject to laws at all, deterministic or statistical, it is subject to physically-stateable laws. If feelings are immaterial and aphysical, they do not interfere with the material or physical cosmos, nor is any exception to otherwise valid laws required on their account.

Yet to say that what the physical account of the world leaves out is unimportant for physical science is not at all to say that what the physical account
of the world leaves out is unimportant. On the contrary, what is left out of the physical account of the world is thinking and above all feeling, and the realm of feeling is arguably the locus of all non-instrumental value, without which, to paraphrase a remark attributed by Herbert Feigl to Albert Einstein, the physical cosmos would be a dunghill. (Galen Strawson 2018 conjectures that the original German term, which Feigl does not give, was rather stronger than “dunghill.”)

Now if all this is granted, what is there left of the mind-body problem? What more is there to say about the relation of the mental to the physical? Has any issue of importance been overlooked? Well, from the beginning we have been concerned in this study with mentality in the sense of human mentality, and therewith feelings in the sense of human feelings. But whatever non-human feelings there may be, they are of value as well, and so, despite the fact that I have largely left it out of account up to this point, and will largely continue to do so, it is a question of importance what non-human feelings there are. Of moral importance in particular is the question which animals can feel pain — or rather, as Jeremy Bentham more correctly put it, which animals can experience suffering — and to what degree. There is a quite considerable literature on the topic, which now has a whole on-line journal of its own, Animal Sentience (founded by Harnad and inaugurated with his position paper Harnad 2016) whose first issue was devoted to the question of pain in fish. I will, accordingly, limit myself to a very few remarks here.

In a very brief look earlier on at the issue of other minds, I noted some tendency for children to engage in what adults would regard as overgeneralization, and then through critical reflection, or simply under adult influence, to retract some spontaneous generalizations later. Speaking in very general terms, there are two routes to critical retraction of spontaneous generalization in attributing feeling to animals: looking more closely at behavior; and looking beneath external behavior to internal mechanisms. The latter route only becomes
available, of course, when one has some substantial amount of knowledge about the internal mechanisms at work in the one case where there is no serious disagreement among sane persons as to whether feeling and thought are present, namely, our own, human case. The difference between the routes can be illustrated by a hypothetical case suggested by the finger-on-the-hot-stove example.

So suppose we have some type of small aquatic worm, swimming in an aquarium with bristles in it here and there. Those on one side have no particular effect, and the stimulus of being touched by them produces no detectable response. Those on the other side administer some kind of electric shock, and if the worm is touched by one, it will very quickly twist away the body segment in contact with it. The naive (and quite possibly entirely correct) reaction is to suppose that the bristles on the second side hurt the worm.

One way one might be brought to doubt this initial reaction would be by a closer, more extended look at the worms’ behavior. Suppose it were found, for instance, that the worms, despite sharply withdrawing from the bristles of the second kind when touched by them, never come to avoid the side of the aquarium where those bristles are present, as surely eels would if we scaled up the whole experiment and substituted a more intelligent organism. The absence of learning-behavior might lead to questioning whether any pain was really felt.

Another way one might be brought to doubt the initial reaction would be by internal, anatomical examination of the worms. I mean more specifically, anatomical examination comparing what goes on in the worm when it touches an electrified bristle and what goes on in a human being when one touches a hot stove. Suppose it were found that, though the worm has ganglia like those of our spinal cords, there is nothing at the anterior or head end comparable to our brains, and that the withdrawal upon being shocked only involves nerves connecting the skin and muscles of the segment impacted to the nearest ganglion, and is not
accompanied by any movement of an impulse up towards the head. The absence of any involvement of a brain might not only explain the absence of learning but also lead to serious questioning whether any pain was really felt. (Inversely, discovery that some sea organism we had thought was a plant is in fact an animal with nervous tissue might lead to reclassifying as sentient what had been taken for insentient.)

The advance of science, both ethological close study of animal behavior, and physiological study of animal nervous systems, can be expected to provide more and more pertinent data. But it cannot in the end tell us how to evaluate the data provided, or what ultimate conclusion to draw. How much weight should be attached to the kind of phenomena a scientist taking a behavioral or functional approach would note? And how much to neurophysiology — for instance, to the presence or absence, or scarcity or abundance, of C-fibers, or of structures homologous to those found in the human amygdala? What the journal issue on piscine pain and suffering most clearly shows is that there is at present no consensus as to how to weigh different bits of information, with the result that a significant part of the discussion on moral aspects of the issue has to be devoted to the question of what is the appropriate course of action to take when we remain doubtful and uncertain about whether pain and suffering does or does not occur in a given case.

I bring all this up, not because I have ever pretended to have any new insight to offer on the issue of animal sentience — recognition that evolutionary considerations are irrelevant may save us from fallacious inferences, but certainly doesn’t make our questions any easier to answer — but merely to contrast that question with others that still remain before us even when interactionism and denialism have both been rejected, leaving us with just the division between lacunist physicalism and epiphenomenalist dualism. The fish are out there,
swimming around, and the question whether they are hurting quickly becomes, for those who engage with it, a real question of moral import. By contrast, the metaphysical issue left once physicalism and dualism have been pared down to their lacunist and epiphenomenalist versions can for certain temperaments rather easily come to seem artificial and unreal in comparison.

The metaphysical issue remaining may be put this way: granted that there is something important left out of the physical account of the world, is there some thing left out? Not some substance, some kind of stuff — that much seems agreed — but perhaps a type or token of something else: some property, or event, or state, or fact. There is room for a question here because of a special feature of the logic of such expressions as “there is something,” which Friederike Moltmann (2013) has called special quantifiers. A simple example may illustrate the logical peculiarities of such expressions as compared with more ordinary quantifiers such as “there is someone” or “there is person,” or any instance of “there is some S” where S is a sortal expression.

Jack, leaving the apartment, calls out, “I’m heading to the bank to check about the loan.” Jill, in an adjoining room, replies, “There’s something you’re forgetting.” Jack inquires, “What?” At this point Jill may make either of two logically quite different types of response. One the one hand, she may say, “Your umbrella. It will be raining cats and dogs by the time you’re done at the bank.” In this case, her answer is a noun phrase, and as such at least purports to denote some thing. (There may, of course, be no such unique thing as “Jack’s umbrella”; but even if the umbrella he used to have has recently become lost or broken, or he recently acquired a second umbrella, still “Jack’s umbrella” at least purports to denote some specific umbrella.) On the other hand, she may say, “That it’s a holiday. The bank will be closed.” In this case, her answer is a that-clause and need not be regarded as even purporting to denote any thing. (Some theorists may hold
that a that-clause denotes a kind of thing called a “proposition,” but that is a contentious metaphysical hypothesis, and one not at all *obviously* true.)

The lacunist physicalist grants that there is something the physical account of the world leaves out, but denies that there is any *thing* that it leaves out. There is something left out by the physical description of a human subject with normal eyes and nerves and brain being exposed to magnetic radiation of wavelength around 700 nanometers, namely, what is expressed by the that-clause “that seeing red is like *this*, and not like *that*;” but the physicalist denies there is a property or event or state or fact distinct from any mentioned or alluded to in the physical description. The quintessentially metaphysical question whether or not there is any such thing is what remains of the mind-body problem for those who have followed the dialectic as I have described it down to this point. To be of anti-metaphysical temperament is to find such a question less compellingly real than the question whether, and if so what, fish feel.

Over time, the philosophical community has seesawed between periods of high metaphysical speculation, seeking to get behind all merely human representations to how ultimate reality is in itself, independently of us and our concerns, and flattering itself that it has succeeded in this quest, and periods of doubt and disbelief. For there is a tendency for metaphysical systems to become more and more elaborate, passing from classical in their architecture to mannerist and baroque and rococo, until they collapse of their own weight and a reaction sets in, as happened towards the middle of the nineteenth century with German Idealism, and then in the early twentieth century with its revival, British Idealism.

The analytic tradition in philosophy arose as a reaction on the part of Russell and others against idealists in the generation of their teachers. One school within the analytic tradition, logical positivism as represented by Carnap (1950), being attached to a principle connecting meaning with method of verification, concluded
that metaphysics is meaningless, and its problems pseudo-problems, on the ground, roughly speaking, that there was not enough agreement among metaphysicians as to what would count as verification of this or that metaphysical thesis to give such theses a meaning. That is why, according to Carnap, debate goes on and on without ever reaching resolution.

Another school, ordinary language philosophy as represented by P. F. Strawson, more interestingly held that there is room for something that might inherit the label “metaphysics” even after repudiation of the characteristic ambitions of the kind of speculation indulged in by F. H. Bradley or G. W. F. Hegel. For Strawson held there was room for an inquiry he called descriptive metaphysics (not unrelated to the enterprise engaged in by Moltmann today), concerned with the analysis of the central concepts of our commonsense thought, without regard to its relation to any ultimate reality of the kind he and his contemporary analytic philosophers of many schools did not take to be accessible to fruitful inquiry. While the gravest epistemological worries about the project of speculative metaphysics were felt, the methodology of descriptive metaphysics was not supposed to be fundamentally different from that of philosophical analysis at large, or for that matter, of lexicography.

But the conclusions of descriptive metaphysics are often at variance with views enjoying a widespread acceptance in speculative metaphysics, which Strawson called “revisionary” metaphysics. This is perhaps most easily seen in the case of facts. If we look at the descriptive metaphysics of facts, which is to say, at the ordinary usage of the word “fact,” we find a conception that makes facts both abundant and fine-grained, while many speculative metaphysicians may take them to be sparse and/or coarse-grained.

Facts as descriptive metaphysics would have to describe them are abundant in the sense that whenever I may affirm that \( P \) I may with equal right begin to speak
of “the fact that $P$.” If there are no unicorns, then ordinary language immediately allows this assertion to be pleonastically expanded to the assertion, “It is a fact that there are no unicorns.” One may even say, “The fact that there are no unicorns explains the fact that those who go out hunting unicorns never come back bringing any with them.” Quite a few speculative metaphysicians, especially though not exclusively those involved in debates about so-called truth-makers and the traditional correspondence theory of truth, maintain by contrast that there are no negative existential facts, no facts about the nonexistence of kinds of things, unicorns included. (Many deny also the existence of disjunctive facts.)

As for the gauge or grain of facts, the matter is perhaps best approached through consideration of awareness of facts, but here a slight caution is in order at the outset. It must be pointed out that a subject may, if not fully linguistically competent in a given language, be aware that $P$ without being disposed to answer affirmatively to the question whether $P$ when put in that language. To borrow half an example from Wittgenstein, a dog may well be aware that its master is at the door, but if you ask the dog, “Is your master at the door?” it is unlikely to bark “Yes.” Or as I recall happening in a B-movie or C-movie comedy years ago, a boxer may not answer affirmatively the question, “Are you a pugilist?” and if mistaken enough about what the word “pugilist” means — confusing it with some other word beginning with the same letter — may even react in a violently negative way. But questions of linguistic competence aside, one who may be said to be aware that $P$ may in ordinary language equally well be said to be aware of the fact that $P$, and similarly for unawareness.

Facts as descriptive metaphysics would have to describe them are fine-grained in the sense that, except where $P$ and $Q$ are (to the linguistically competent) pretty obviously equivalent, the fact that $P$ is a distinct fact from the fact that $Q$. For unless $P$ and $Q$ are pretty obviously equivalent, some person might be in a state
ordinary language would describe as one of being aware that \( P \) and unaware that \( Q \) (or vice versa). And then ordinary language immediately allows this assertion to be pleonastically expanded to the assertion that this person “is aware of the fact that \( P \) but not of the fact that \( Q \).” But if the fact that \( P \) is one of which this person is aware, and the fact that \( Q \) is one of which this person is unaware, then these must be two distinct facts.

Speculative metaphysicians, by contrast, often claim that facts are coarse-grained, and that in many cases where ordinary language and commonsense thought would make a distinction in the way just described, anyone aware of the fact that \( P \) must be *ipso facto* aware of the fact that \( Q \), since these are quite literally one and the very same fact. For instance, someone as little disposed as Homer to answer affirmatively the question “Is Hesperus the same as Phosphorus?” even in Greek translation, will be claimed by some theorists, contrary to the view of Kripke, to be nonetheless well aware of the fact that Hesperus is Phosphorus, on the grounds that it is quite literally one and the very same fact as the fact that Hesperus is Hesperus. This difference is directly relevant to the dualist-physicalist debate in cases where \( P \) is about a conscious state while \( Q \) is about some physical correlate.

To give an example from personal experience, some years ago I had for ten or minutes or so a striking hallucination, as of water running down the walls of a small room where I happened to be. There had previously been problems with leakages from the plumbing above this particular room, and I at first believed that the plumbing problems had recurred, until I touched the wall and found it bone dry. Then I did not know what to think.

The experience was not disturbing enough to send me to a doctor, but when I did in the normal course of events have to go to various doctors, I did ask about it, but received only funny looks and no enlightenment until I mentioned it to my
ophthalmologist, Dr Samuel Liu. He was able, after asking a very few apt questions, to give a probable diagnosis of optical migraine. Prior to consulting with Dr Liu, I had been aware of having had a strange experience, but not of what my kind of experience is called; after consultation, I was aware of having had a strange experience of the kind called an “optical migraine.” I was not, however, and still am not, aware of what goes on in the nerves or brain during such an event.

Dr Liu, who presumably has knowledge of the neurology of these events, could therefore say that “JB is aware of the fact that he has had such-and-such an experience but unaware of the fact that thus-and-so went on in his nervous system.” And from the point of view of descriptive metaphysics, the conclusion now follows that the physicalistically-indescribable fact that I have had such-and-such an experience is distinct from its correlated neural fact, that thus-and-so went on in my nervous system, or from any physicalistically-describable fact. I leave it as an exercise to the reader to think through how what I have said about facts might apply to states or events or properties.

The long and short of it is that from the standpoint of descriptive metaphysics, neuralism is wrong, and for similar reasons psycho-functionalism is wrong, and so physicalism is wrong, and dualism is right, and supervenience fails.

6.2 Modal Knowledge

Dorothy grew up in Kansas, where everything is gray, until one day a cyclone picked her up and set her down in a technicolor land where she encountered a rainbow of hues, and she was able at last to say to herself, “So this is what emerald green looks like!” or “So this is what it is like to be seeing emerald green!” According to Jackson’s former view, and the line he originally took in the case of his Mary, even if Dorothy back home had learned everything there is to be learned about color and color vision from black and white books and black and
white newsreels, still she would now be learning a new fact. And I have just argued that from the standpoint of descriptive metaphysics, Jackson would have been right. For those who think the only good metaphysics is descriptive metaphysics, that settles the question. For speculative metaphysicians it does not.

A speculative metaphysician might want to claim that though Dorothy may, looking at an appropriate sample of an emerald, come out with the kind of utterances I have suggested, still there is, owing to the sparsity and/or coarse-grainedness of facts, no new fact involved beyond the physicalistically-describable, neural or functional, facts of which she knew all along from her book-reading and newsreel-viewing back on the gray prairies of the American Midwest. Lewis, a whiz of speculative metaphysician if ever a speculative metaphysician there was, took just such a line in the case of Mary.

It would be less misleading if speculative metaphysicians used a different expression, say “factoids” or “FACTS” for their coarse-grained items, leaving the ordinary word “facts” to the descriptive metaphysician for facts of the ordinary (or ordinary-language) fine-grained kind. As we have seen, if there are such thing as facts in this latter sense, then they include mental facts that are not physical facts. Thus physicalists who want to maintain that every mental thing is a physical thing seem to need to deny that there are such things as facts in the ordinary sense, though it is open to them to accept FACTS in some extraordinary speculative- or revisionary-metaphysical sense.

Alternatively, they could revisionistically redefine “physicalism,” not as the doctrine that every mental thing, facts included, is (identical with or supervenient on) a physical thing, but only as the doctrine that every mental THING, FACTS but not facts included, is (identical with or supervenient on) a physical THING. It could be granted that non-physical, mental descriptive-metaphysical facts exist, but claimed that they are immaterial, not in the obvious sense, but in the sense of
“immaterial” illustrated by the quotation from Ambrose Bierce that appears as the epigraph to this book: “unimportant.” (For Bierce, only what is “immaterial” in the obvious sense, thought and feeling and vision, is “material” in the sense of having importance.) The so-called hyperintensional jargon of “essence” and “fundamentality” and “grounding,” with its attendant apparatus, might supply a technical vocabulary in which to express this line, and distinguish things that are mere creatures of human convention from THINGS populating a realm of reality behind mere human linguistic or other representation.

A shift like this has taken place in certain quarters in philosophy of mathematics. Philosophers who a generation ago would have called themselves “nominalists,” and denied that numbers and sets and the like exist, may now still call themselves “nominalists,” while conceding that numbers and sets and the like do after all exist, though insisting that they are somehow unimportant or second-rate things, not important or first-rate THINGS (or alternatively, that they merely exist and do not EXIST). A similar redefinition of the mind-body problem and “physicalism” would be involved in taking the line on the mental that I have been alluding to.

Such verbal, redefinitional manœuvres cannot affect the conclusions reached before even taking up the question of facts, such as the conclusion that there can be no evolutionary explanation of consciousness. But redefinitional manœuvres do have a capacity to kick up dust and obscure our view of significant conclusions, and suspicious dualists will be leery of such manœuvres and on the look-out for abracadabra and on guard against hocus-pocus here.

So for simplicity let me not pursue the line I have been sketching, and rather retain the now perhaps somewhat old-fashioned way of describing matters, on which the speculative metaphysician simply denies that there is any new fact for Mary or Dorothy to learn, rather than saying there is a new fact, but not an
important one, where “important” here would be a place-holder for some fancier jargon. So on this simpler way of taking matters we have the descriptive metaphysician committed to dualism, and a speculative metaphysician committed to a physicalism that is based on an understanding of “fact” or “state” or “event” or “property” or whatever differing from that implicit in ordinary language.

Since the supervenience thesis marks the boundary between dualism and physicalism, to the degree that descriptive metaphysics is associated with dualism, and speculative metaphysics with physicalism, the one tends towards a rejection, and the other towards an acceptance, of supervenience. The contrast between different metaphysical projects and the conclusions about the modal issue of supervenience most congenial to them can be found associated — I am speaking here of a psychological tendency, not a logical inevitability — with an opposition between two different orientations towards modality itself. It is on this difference of orientation that I would like to focus now, bringing forward a matter that has for some time been lurking in the background, before wrapping up.

Let me begin with what, however little it may be of concern to new-wave hyperintensional thinking, it is hard to deny has been for much of the history of modern philosophy one of the chief and perhaps the chief mystery about necessity and possibility. I mean the problem of our knowledge of necessity. Here is how the question arises. We think that when things are some way, in some cases they could have been otherwise, and in other cases they couldn't. And we not only think that there is such a modal distinction, but also think that we know some examples. For instance, though as of this writing there has been no female U. S. president, there could have been, whereas the number 29 not only has no nontrivial divisors, but couldn't have had any. Thus we may consider the counterfactual, “What if Hilary Clinton had been president during the coronavirus pandemic in 2020?” but won’t consider the counterfactual, “What if 5 had been a divisor of 29?” And what is
called the question of modal epistemology is just this: How do we know that the examples in question are examples of that of which they are supposed to be examples?

And why should this question be considered a difficult problem, a kind of mystery? Well, that is because, on the one hand, when we ask about most other items of purported knowledge how it is we can know them, sense-experience seems to be the source, or anyhow a source of our knowledge; but, on the other hand, sense-experience seems able only to provide knowledge about what is or isn’t, not what could or couldn’t have been. How do we bridge the gap between “is” and “could”?

I have already followed Kripke in quoting the classic statement of problem from Kant, his remark about what experience can and cannot teach us. Kant’s formulation leaves open that experience can teach us that a necessary truth is true; what it is supposed to be unable to teach us is that it is necessary. For example, experience with what they call “manipulatives” in pre-school, trying to arrange 29 blocks in various patterns, may convince a child that 29 is a prime number, one that has no nontrivial factorization. But granted that it doesn't actually have one, how do we know that it couldn't possibly have had one, and is necessarily a prime number? How can experience teach us that extra fact?

The problem becomes more vivid if one adopts the language that was once used by Leibniz, and much later repopularized by Kripke in his famous pre-N&N technical work on modal logic, the usage according to which the necessary is characterized as “that which is true in all possible worlds.” In these terms the problem is that the senses only show us this world — look around you — the world we live in, the “actual” world as it is called, whereas when we claim to know about what could or couldn’t have been, we are claiming knowledge of what is going on in some or all other worlds. For that kind of knowledge, it seems, we
would need a kind of sixth sense, or extrasensory perception, or nonperceptual mode of apprehension, to see beyond the world in which we live to these various other worlds.

Skipping over the history of the problem between Kant and the twentieth-century, by the time we come to Carnap or A. J. Ayer, representing logical positivism, and P. F. Strawson, representing ordinary language philosophy, we have the necessary and the apriori and the analytic identified with each other and with what follows from semantic rules or linguistic conventions. On this picture, we know that five isn't a divisor of 29 in any possible world, not because we are able to engage in telepathic remote viewing of other possible worlds while remaining in this one, but rather because, just as we have implicitly or unconsciously learned the rules or conventions that tell us no one who is married may properly be described as a “bachelor,” so also we have implicitly or unconsciously learned rules or conventions that tell us, or from which it follows, that one may not describe any world as a world in which 5 is a “divisor” of 29.

For instance, a world in which whenever one lays out five rows of six blocks and then counts the total, one finds 29, is not properly described as a world in which five times six is 29. Rather it is a world where, by some curious physical principle, whenever 30 blocks are brought together in five rows of six, one of them disappears; or where, by some curious psychological principle, whenever we try to count 30 blocks arranged in five rows of six, we always miss one; or something of the sort. This, I say, is the sort of picture that had become the received wisdom in philosophy departments in the English-speaking world by the middle decades of the last century, the era of Ayer and P. F. Strawson, who though they disagreed with each other across a whole range of issues, both put forward a picture quite like the one I have just drawn (in Ayer 1936 and P. F. Strawson 1952).

Both would sometimes go on to speak as if semantic rules or linguistic
conventions were the source not only of our knowledge of modality, but of modality itself, and go on further to speak of the source of modality lying in ourselves. For while individually, as children growing up in a linguistic community, or foreigners seeking to enter one, we must internalize the pre-existing implicit rules of the communal language, collectively, as a speech community, we do not so much learn as create the language with its rules. If the origin of modality, of necessity and its distinction from contingency, lies in language, it therefore lies in a creation of ours, and so in us.

Thus Ayer, in his chapter on the apriori says this:

The principles of logic and mathematics are true universally simply because we never allow them to be anything else. (p. 41)

And Strawson early on, speaking of why some pairs of predicates are incompatible, says this:

It is we, the makers of language, who make predicates incompatible … It is we who decide where the boundaries are to be drawn. (p. 5)

The common element here is the pronoun “we.” The picture is that the necessary, identified with the apriori, identified with the analytic, is made so by us, that we are the source. Such was the climate of opinion before January, 1970, when Kripke delivered his Princeton lectures that became N&N.

Since then it has become clear that there are several things wrong with the conclusion that we, the makers of language, are the originators of necessity, and the line of thought from Kant onwards that eventually led to this sort of conclusion. I won’t try to say in advance just how many main things are wrong with this line of thought, lest I end up like Cardinal Ximenes in the Monty Python sketch about the
Spanish Inquisition, and find myself repeatedly having to revise my estimate upwards. I'll just mention several difficulties one by one and you can count them up afterwards if you wish.

To begin with, there is a difficulty discussed by Kripke in his addenda, in a passage about the transition from the necessary to the apriori that I alluded to earlier, Kant’s seeming transition from “Experience isn’t sufficient to teach us that a thing couldn’t have been otherwise,” to “Experience isn’t needed to teach us that a thing couldn’t have been otherwise.” Strawson notoriously called one of Kant’s arguments a “non-sequitur of numbing grossness,” and whether or not the argument to which he applied that label — if memory serves, it was the second analogy — deserved it, the argument from “Experience isn’t sufficient…” to “Experience isn’t needed…” certainly does. It’s like arguing from “Sunshine isn’t sufficient to grow corn (rainfall is also needed)” to “sunshine isn’t needed to grow corn (rainfall is sufficient).”

The achievement that made Kripke’s lectures an overnight sensation — I can remember when I was a graduate student in the Group in Logic at Berkeley my dissertation supervisor, the late Jack Silver, a technical mathematical logician who never published a word on philosophy, telling me about Kripke’s views with an air of excitement shortly about the publication of the lectures, though he kept me busy enough with other work that I did not find time to read them until I took up a post-doc at Madison a couple of years later — was not just discovering a logical lapse in a famous philosopher, a kind of discovery that is, alas, all too common. Rather, it consisted in producing plausible examples of aposteriori necessities about everyone and everything from the genealogy of the Queen to the taxonomic position of Princeton’s mascot, the tiger.

Less noticed was the remark in his addenda, to which I have also already alluded, to the effect that behind each of these examples there is some apriori
principle according to which whatever is so about certain matters is so of necessity. On this picture, we know a priori that whatever the facts are about whether one number divides another, they couldn’t have been otherwise. Experience with manipulatives may teach us that five doesn’t divide 29, and combining what experience thus teaches us with our background apriori knowledge, we can conclude that five couldn’t have divided 29, and similarly with other examples.

Kripke describes this a priori knowledge in the background as “discoverable by philosophical analysis,” but does not quite explicitly call it “analytic.” Still, it seems to be at least compatible with what he says to hold that our knowledge of necessity reduces to a combination of a posteriori knowledge gained mainly through sense-experience with analytic knowledge of semantic rules or linguistic conventions or whatever.

In that case the formulation that our knowledge of necessity derives from our knowledge of linguistic rules would be defensible if understood as meaning only this, that though our knowledge of the truth of a necessary truth may depend on sense-experience, our knowledge of the necessity of that truth, given that it is true, derives from knowledge of linguistic rules. The main point is that the epistemology of modality is still demystified: there is no need for ESP. And perhaps that is what the logical positivists and ordinary language philosophers really cared about.

Unfortunately for them, Quine (1936) had already found some serious difficulties with the notion of truth by convention, and Wittgenstein had raised puzzles in cryptic paragraphs about rules, from which Kripke (1982) extracted and developed a so-called skeptical paradox about meaning in general. For present purposes it will suffice to recall Quine’s version, seen from Kripke’s angle.

I wrote above casually about necessity being a matter of, or at least following from, rules or conventions, but the gap between “being a matter of” and “following from,” looked at from a certain perspective, becomes a yawning chasm. There are
infinitely many necessary truths, certainly if we count mathematical truths among them, but there has only been time for human beings collectively to establish and individually to absorb finitely many. In most cases it must be a matter of a necessary truth’s merely following from something more basic. But then what kind of fact is it that it thus follows? It would seem to be a necessary fact that has not been traced to rules or conventions.

This is a large problem, so large that I will not attempt to deal with it (and Kripke himself never published a proposed solution he could endorse in his own person), but simply assume that it can be solved somehow. If so, then the point we have reached is this, that while Kripke has at least taught us that we must be very careful about how we understand such a formulation as “Our knowledge of necessity derives from knowledge of linguistic rules,” he has not really shown, or even claimed very strenuously, that such a formulation is wholly off the mark.

There remains, however, to be evaluated the step from a claim about the origin of knowledge of necessity to a claim about the origin of necessity itself, such as seems to be advanced or insinuated in the quotations from Ayer and Strawson produced earlier, the claim that necessity itself originates from our linguistic rules — presumably rules about using modal auxiliary verbs — and beyond that the claim that necessity originates in us, the makers and users of language.

6.3 Modality’s Origin: Sacred or Profane?

One immediate difficulty with this last conclusion is that it, along with all talk about the “origin” of necessity, is simply nonsense: the problem of the origin of necessity is not a mystery but a muddle. For if one takes the word “origin” at all literally, then to speak of the origin of necessity is to speak of a time when necessity came into being, and before which there was none. And that certainly seems to be nonsense. The number 29 not only is today but always has been
necessarily prime. It is absurd to speak as if before we were on the scene to adopt any linguistic rules to prevent it, 29 was free to be composite if it wished.

It is equally absurd to say that if we had adopted different linguistic rules, 29 would have been composite. To be sure, if we had adopted different rules, the orthographic or phonetic sequence of symbols or sounds “29 is composite” might have been true in the language established by those hypothetical rules. For we might have adopted base thirteen instead of base ten, in which case “29” would have denoted 35, which is composite. But even then, “29 is composite” would have been false under our actual rules, and 29 would not have been composite, but prime.

So in speaking of ourselves as the originators of modality, the philosophers of the mid-twentieth-century will have been speaking nonsense if they were speaking literally. But surely they, or the best of them, must be credited with having had the wit to see this, and so we must take them to be speaking figuratively, even if they make no particularly strenuous efforts to warn against literalism or to spell out in terms that can be taken literally what they express figuratively. Let us grant this, and let us allow ourselves to go on speaking figuratively for a bit. Even so, the formulation according to which we are the creators of necessity is open to at least one further challenge.

Even granting that what is needed to get from the kind of knowledge sense-experience can give us to knowledge of the necessity of necessary truths is knowledge of linguistic rules, and even granting also that despite skeptical paradoxes we do have the requisite kind of knowledge of linguistic rules, still it is premature to conclude, even speaking figuratively, that necessity itself derives from linguistic rules, and so from us.

For another interpretation of the situation remains possible, dismissing modality in a sense tied to the grammar of ordinary language in favor of MODALITY
in another and more speculative or transcendent sense, making a distinction roughly parallel to that between facts in a sense tied to the grammar of ordinary language in favor of FACTS in another and more speculative or transcendent sense. Logical positivists and ordinary language philosophers have given no real argument against this rival interpretation, but seem merely to have assumed, such being the intellectual climate of their days, that the rival view, to the extent they were aware of it, is too metaphysical, in a pejorative sense, to be taken seriously by analytic philosophers.

The real problem or mystery of the so-called origin of necessity lies in the clash between these two interpretations, and my next goal must be to try to bring out what the conflict here consists in. So let me now try to describe the two rival views, and after that explain the significance of their rivalry. I will give first the view that people like Ayer and Strawson want to reject.

Here is the first view. Though the event is not mentioned in Genesis, the first thing God said on the first day of creation was “Let there be necessity.” And there was necessity. And God saw necessity, that it was good, and God divided necessity from contingency — and, having done all this in the dark, only then took up the question of light. Several days later, Adam and Eve, in addition to introducing names for the animals, in a short break between the birds and the fish introduced modal auxiliary verbs and rules for their use, rules according to which it can be said of some things that they “could have been otherwise,” and of other things that they “could not have been otherwise.” In so doing they were merely putting labels on a distinction that was no more their creation than were the birds of the air or fishes of the sea that they were naming.

And here is the rival view. The failure of Genesis to mention any command “Let there be necessity,” is to be explained simply by the fact that no such command was issued. We have no reason to suppose that the language in which
God speaks to the angels contains modal auxiliary verbs. Sometime after the Tower of Babel, some of the seventy or seventy-two nations found that their purposes would be better served by introducing into their languages certain modal auxiliary verbs, and fixing certain rules for their use. When we say that such-and-such is necessary while such-and-such is contingent, we are applying such rules, rules that are products of human intelligence, not divine wisdom.

I have been allowing myself here to use theological language in describing the distinction I want to draw. That might have been the natural way for seventeenth or eighteenth century philosophers to discuss the matter. For many today, such language cannot be literally accepted, and if it is only taken metaphorically, then I have not really done better than those who speak figuratively and frame the question as that of whether the so-called origin of necessity lies outside us or within us. So let me drop the theological language, and try to describe without it the division between two views that I have in mind.

Here, then, is a second take on first view. Ultimately reality as it is in itself, independently of our attempts to conceptualize and comprehend it, contains both facts about what is, and superfacts about what not only is but had to have been. Our modal usages, for instance, the distinction between the simple indicative “is” and the construction “had to have been,” simply reflect this fundamental distinction in the world, a distinction that is and from the beginning always was there, independently of us and our concerns.

And here is the second view, take two. We have reasons, connected with our various purposes in life, to use certain words, including “would” and “might,” in certain ways, and thereby to make certain distinctions. The distinction between those things in the world that we say “would have been no matter what” and those we say “might have failed to be if only” is a projection of the distinctions made in our language. Our saying there were necessities there before us must be called a
retroactive application to the prehuman world of a way of speaking invented and created by human beings in order to solve human problems.

Well, that's my second try. With it I have gotten rid of theology, but unfortunately I have not gotten rid of all metaphors. The key remaining metaphor is the optical one: reflection versus projection. Perhaps I should give up the attempt to get rid of all metaphors, and admit that the two views I am discussing should be labeled not so much theses or doctrines as attitudes or orientations: a stance that finds the “reflection” metaphor congenial, and the stance that finds the “projection” metaphor congenial. But let me try a third time to describe the distinction between the two outlooks in literal terms, avoiding optics as well as theology.

To begin with, both sides grant that there is a correspondence or parallelism between two lists of items. On the one hand, there are truths about the contrast between what is necessary and what is contingent, for instance, between 29 being a prime number and 29 being the number of years it takes for Saturn to orbit the sun. On the other hand, there are truths about our usage of modal auxiliary verbs such as “would” and “might,” for instance, the fact that we have no use for questions of the form “Would 29 still have been a prime number if such-and-such?” but may have use for questions of the form “Would 29 still have been the number of years it takes for Saturn to orbit the sun if such-and-such?” The difference between the sides concerns the order of explanation of the relation between the two parallel ranges of truths.

And what do I mean by that? Well, both sides grant that “29 is necessarily prime,” for instance, is a proper thing to say, but they differ in the explanation why it is a proper thing to say. Asked why, the first or reflectionist side will say that ultimately it is simply because 29 is necessarily prime. That makes it proper to say that it is necessarily prime, to assert the proposition that it is necessarily prime, and
since the sentence “29 is necessarily prime” expresses that proposition, it is proper
to utter that sentence assertively. The second or projectionist side will say instead
that there is a rule of our language according to which, or from which it follows
that “29 is necessarily prime” is a proper thing to say, and that is why it is a proper
thing to say. Note that the adherents of the second view need not deny that 29 is
necessarily prime. On the contrary, having said that the sentence “29 is necessarily
prime” is, per rules of our language, a proper thing to say, they will go on to say it.

Nor need the adherents of the first view deny that recognition of the propriety
of saying “29 is necessarily prime” is somehow incorporated into the system of
rules of our language. The adherents of the first view need not even deny that
proximately, as individuals, we learn that “29 is necessarily prime” is a proper
thing to say by picking up the pertinent rule in the course of learning our language.
But the adherents of the first view will maintain that the rule itself is only proper
because collectively, as the creators of the language, we or our remote answers
have, in setting up the rule, managed to achieve correspondence with a pre-existing
fact, or rather, a pre-existing superfact, the superfact that 29 is necessarily prime.
The difference between the two views is, as I said, in the order of explanation.

6.4  Modal Knowledge, bis

The two rival views of modality, projectionist and reflectionist as I have
termed them, bring with them, or ought to bring with them, different agendas and
priorities for developing a theoretical account of modal distinctions. On the
projectionist view, the first task must simply be to characterize accurately the rules
for the use of distinctively modal talk implicit in ordinary language. That this is no
easy matter can be seen from the circumstance that, though countless systems of
modal logic have been put forward and their formal or technical sides thoroughly
explored, especially since Kripke’s development of a fruitful model theory, there is
still no agreement as to which if any of these systems best represents the logic of [metaphysical] modality as embodied in ordinary language. Kripke's model theory is customarily called a “semantics,” but any connection with semantics in the original sense of a theory about meaning is tenuous.

That the descriptive task is difficult should be no surprise. For any task involving tracing patterns in commonsense thinking presents difficulties and encounters paradoxes. An even more famous quotation from Kant than the one about what experience teaches us is the famous line from his *Idea for a Universal History*, according to which “From the crooked timber of humanity, no straight thing was ever built.” It is a feature of anything created by us, including our systems of linguistic rules and conventions as much as anything else, that it should be full of irregularities and anomalies. Our rules and conventions may be called “lawlike,” not in the sense in which philosophers generally use that term (and I myself have used it so far), but in the contrary sense of being literally like the law.

And what is the law like? It is a hodge-podge of elements derived from disparate sources: in the case of American law, ancestral customs of the Angles and Saxons and surviving fragments of Roman law, principles that came into use in courts of law and principles that came into use in courts of equity, bodies of statute law and masses of case law, all several times at different periods half-overhauled and semi-recodified by reform movements that achieved partial but not total success before running out of steam. Hardly any principle is truly universal; rather, one has general rules with particular exceptions. Something as messy as the law is what a completely successful description of our existing patterns of modal speech and thought can be expected to reveal.

I mentioned earlier the fact that those attempting to justify the continued use of teleological language of “function” and the like in biology in the wake of natural selection will often try to explain talk of the “function” of an organ as being about
whatever it is about the organ that makes its presence conducive to survival and production of fertile offspring. Were the task of accurate description of our conventions pertaining to modal distinctions ever completed, a second task would then loom for the projectionist, that of determining the function of such rules and conventions in the sense of “function” just indicated. What if anything do we accomplish by making modal distinctions at all, and how if at all do we accomplish it better with the system of rules and conventions we have implicitly adopted than we might have done had we adopted a different system? Addressing something like this question ought to be item number two on the projectionist’s to-do list.

To the reflectionist, the situation must look quite different. The true goal is to describe the ultimate reality that our rules and conventions reflect. The practitioners of hyperintensional metaphysics, in particular, are engaged in developing their own account of the origin of modality, in terms of essence and fundamentality and grounding. (It seems usually to be put forward as if opposing or ignoring views emphasizing the role of convention, though the distinction made so much of in this tradition, between genuinely essential and merely necessary properties of an entity, does have a parallel in a distinction between what is genuinely part of a convention and what is merely a consequence.)

Given what crooked timber we are, the goal of describing ultimate reality is not likely to be accomplishable simply by examining its reflection in us and our language, which is only too likely to resemble the image in a fun-house mirror more than an image in a reflecting telescope. Another methodology seems needed, and arguably more attention to methodological and epistemological issues before pressing forward with substantive questions.

An outsider of a different temperament looking at the reflectionist’s program may well be inclined to think that its first task ought to be a preliminary one of considering closely how we could even hope to find out about an ultimate modal
reality totally independent of us. It is not as if we had faculties responsive to modal reality in a sense of “responsive” implying that if modal reality had been different, our pictures of it would have been correspondingly different. For it makes dubious sense even to speak of the possibility that different things might have been possible than those that actually are possible.

It cannot be said that in the current era of the flourishing growth and vigorous expansion — or metastatic overgrowth and morbid hypertrophy, for those of another temperament — of speculative metaphysics the epistemology of metaphysics in general and modality in particular receives no attention. But for one who is, like myself, more in sympathy with thinkers of the mid-twentieth century, when the pendulum that swings between speculation and skepticism had swung very far in the opposite direction from where it is today, epistemology has not received nearly enough attention. (Certainly those in whom the spirit of Carnap lives on will entertain the gravest of doubts about how we could have knowledge of any distinction between things and THINGS or existence and EXISTENCE, if that is the route the speculative physicalist wants to go.)

Until epistemological issues receive much more attention, one cannot be sure of the right answer to or degree of importance of any question of modal metaphysics, the supervenience issue in philosophy of mind definitely included. One may have one’s suspicions, and I have mine, but however shrewd one may take them to be, they are still only suspicions. That being my view, any conclusion I advance can only be in the nature of something I think can be plausibly suggested, not irrefragably demonstrated. This understood, let me close by enunciating three tentative conclusions or sneaking suspicions about supervenience, using for emphasis and effect exaggerated diction, and leaving it to the reader sprinkle in qualifying phrases where I am most immoderate.

_The supervenience issue involves a pointless detour_. It would be best to
confront directly the question of the identity of a mental state with a neural or functional state. The confrontation may reveal an unbridgeable gap between the intuitions of dualists, who find it obvious that pains are feelings, and like all feelings individuated by who is feeling them and what it is like to be doing so, and those who, on the contrary, find it plausible that pain is in the first instance known by what it does, leaving it open to empirical investigation what sort of thing it is that does that. If so, we will have gone as far as can be gone with the issue, and bringing in modality is not going to get us any further.

*The supervenience issue is a pernicious distraction.* If the real and important issue left open is that of animal pain and suffering, then letting ourselves become entangled in the metaphysics of modality will present not merely an intellectual but a moral hazard.

*The supervenience issue rests on a popular delusion.* I mean the thought currently popular or fashionable again among philosophers that getting behind all merely human representations to ultimate realities is an intelligible and feasible project. This is the thought that leads to dismissing the conclusions of mere descriptive metaphysics, and in particular the conclusion that there are mental facts that are not physical facts, as incompetent, irrelevant, and immaterial. It is the thought that presses philosophers on to meander in a murky, muddy morass of modal metaphysics, where those who have gone before have “found no end, in wandering mazes lost.”

But though I am suggesting that further pursuit of the supervenience issue should be suspended, I certainly believe, and hope the reader will agree, that a good deal has been learned from the issue’s being pursued as far as it has been.
References

Classics from before 1900 by writers from Shakespeare to Nietzsche mentioned in the text by title without date (or simply quoted, like Paradise Lost in the penultimate paragraph) are not discussed in a way that would make differences of editions material, hence are not included among the items listed with bibliographic data below; readers may consult any printed or online version they find conveniently available.

Allen, Colin, and Neal, Jacob, 2020, “Teleological Notions in Biology,” Stanford Encyclopedia of Philosophy,
plato.stanford.edu/archives/spr2020/entries/teleology-biology/
Auburn, David, 2000, Proof, Farar, Straus and Giroux.


www.youtube.com/watch?v=ACkxe_5Ubq8

Elpidorou, Andreas, 2015, “Phenomenal Concepts,” *Oxford Bibliographies Online*


Harnad, Stevan, 2016, “Animal Sentience: The Other-Minds Problem,” *Animal Sentence* 1(1) animalstudiesrepository.org/animsent/vol1/iss1/1/


Papineau, David, 2008, “Kripke’s Proof That We Are All Intuitive Dualists”, core.ac.uk/download/files/99/74068.pdf.


Searle, John, 1980, “Minds, Brains and Programs,” Behavioral and Brain Sciences


Strawson, Galen,


Strawson, P. F., 1952, Introduction to Logical Theory, Methuen.


Watson, John B. 1913, “Psychology as the Behaviorist Views It,” Psychological Review.

Wright, Crispin, 1983, Frege’s Conception of Numbers as Objects, Aberdeen University Press.
Index

abundant vs sparse facts
actuality
affects
alethic modality, see modality, alethic
Allen, Colin
analgesia, congenital
analytic behaviorism, see behaviorism, analytic
analytic functionalism, see functionalism, analytic
analyticity vs syntheticity
animals
anomalous monism
aposteriori necessity
aprioriness vs aposterioriness
Aristotle
Armstrong, David (1926-2014)
Auburn, David
Ayer, Alfred Julius (1910-1989)
basic physical vocabulary
behaviorism
Bennett, Karen
Bentham, Jeremy (1748-1842)
Berkeley, George (1685-1783)
Bierce, Ambrose (1842-1914)
blindness and abilities of the blind,
Block, Ned
Bopp, Thomas (1949-2018)
Boyle, Robert (1627-1691)
Bradley, Francis Herbert (1846-1924)
Broad, Charlie Dunbar (1887-1971)
Buechner, Jeff
Byrne, Alex
Cabanis, Jean Paul Georges (1757-1808)
Carnap, Rudolf (1891-1970)
Carnot, Nicholas Léonard Sadi (1796-1832)
Cartesianism, *see also* Descartes, René
Cartwright, Nancy
causal-role functionalism, *see* functionalism, causal-role
C-fibers
chain of communication picture
Chalmers, David
Churchlands, Patricia and Paul
color
conceivability
consciousness, and self-consciousness
conservation laws
constitution
conventional implicature
correlation principle, *see also* psychophysical correlation
Crick, Francis (1916-2004)
Davidson, Donald (1917-2003)
de Gelder, Beatrice
Darwin, Charles (1809-1882) and Darwinism
Dennett, Daniel
deontic modality, see modality, deontic
Descartes, René (1596-1650)
descriptive metaphysics, see metaphysics, descriptive
descriptivism
Dickens, Charles (1812-1860)
Diderot, Denis (1713-1784)
Disraeli, Benjamin (1804-1881)
Donellan, Keith (1931-2015)
dualism
dynamic modality, see modality, dynamic
eliminativism
Elpidorou, Andreas
Engels, Friedrich (1820-1895)
epiphenomenalism
epistemic arguments against descriptivism
epistemic modality, see modality, epistemic
error, arguments from
error theory
essence, vs accident, and fundamentality and grounding
euphoria
events
evidential modality, see modality, evidential
explanatory gap
facts
Feigl, Herbert (1902-1988)
Fermat, Pierre de (1601-1665), and Fermat's theorem
fictionalism
fine-grained vs coarse-grained facts
fool’s pain
Fourier, Jean-Baptiste Joseph (1768-1830)
free will
Frege, Gottlob (1848-1925)
functionalism, causal-role, analytic, machine-state, psycho-
Galilei, Galileo (1564-1642)
Gell-Mann, Murray (1929-2019) and Weinberg, Steven
Gettier, Edmund
ghosts
Gloucester’s principle
Gödel, Kurt (1906-1978)
Goldbach’s conjecture
Hale, Alan
Harnad, Stevan
Hegel, Georg Wilhelm Friedrich (1770-1831)
Hobbes, Thomas (1588-1679)
Hume, David (1711-1776)
Huxley, Thomas Henry (1825-1895)
hyperintensionality
idealism
identity
ignorance, arguments from
intentional states
interactionism
internalism vs externalism
introspective psychology
Jackson, Frank
James, William (1842-1910)
Jarvik, Robert
Jaynes, Julian (1920-1997)
Johnson, Samuel (1709-1784)
Julius Caesar problem
Kant, Immanuel (1724-1804)
Keller, Helen (1880-1968)
Kim, Jaegwon (1934-2019)
kinematic and dynamic quantities
Kish, Daniel
Kripke, Saul
Leibniz, Gottfried Wilhelm (1646-1716)
Levine, Joseph
Lewis, David Kellogg (1941-2001)
Lichtenberg, Georg (1742-1799)
Liu, Samuel
Locke, John (1632-1704)
Lycan, William
machine-state functionalism, see functionalism, machine-state
Malcolm, Norman (1911-1990)
Mary (fictional character), see Jackson, Frank
materialism, see also physicalism
metaphysics, descriptive, revisionary, speculative
McLaughlin, Brian
Mendeleyev, Dmitri (1834-1907)
mental states
metalinguistic theory of names
metaphysical modality, see modality, metaphysical
modal arguments against descriptivism
modality, alethic, deontic, epistemic, evidential, metaphysical, see also necessity, possibility
Moltmann, Friederike
Monty Python
multiple realizability
Nagel, Thomas
names, proper and common
Neal, Jacob
necessity vs contingency
neuralism
Newton, Isaac (1642-1727)
nominalism
Norton, John
occasionalism
optical migraine
pain and suffering
Palmer, Frank (1922-2019)
panpsychism
Papineau, David
parapsychology
Pascal, Blaise (1623-1662)
pedigree
phenomenal concepts
physicalism, *see also* materialism

possibility, conceptual, *see* conceivability, metaphysical, nomological, suppositional properties

Pryor, James

psycho-behaviorism, *see* behaviorism, psycho-

psycho-functionalism, *see* functionalism, psycho-

psychophysical correlation

psychophysics

Putnam, Hilary (1926-2016)

quantum mechanics

Quine, Willard Van Orman (1908-2000)

realizer, *see* role and realizer

reduction

reference, and sense

reflection and projection

representational theories of mind

rigidity

Robinson, William

robots

role and realizer

Russell, Bertrand (1872-1970)

Ryle, Gilbert (1900-1976)

Salmon, Nathan

Schwarz, Wolfgang

Searle, John

second-order states
sensations
sentience
Silver, Jack (1942-2016)
Skinner, Burrhus Frederic (1904-1990)
Smith, Scott Meadows
Soames, Scott
speculative metaphysics, see metaphysics, speculative
states
Strawson, Galen
Strawson, P. F. (1919-2006)
substance
sunrise and sunset
superfacts
supervenience
Tarski, Alfred (1901-1983)
thought, occurrent and dispositional, conscious and unconscious
tin and zinc
two-dimensionalism
types and tokens
uniformity principle
vitalism
Watson, John Broadus (1878-1958)
Weber’s law
Wigner, Eugene (1902-1995)
Wilberforce, Samuel (1805-1873)
Wiles, Andrew
Wittgenstein, Ludwig (1889-1951)
Wright, Crispin
Zeilberger, Doron
zombies
[≈52K words]