knowledge argument infers from failure of deducibility to difference in facts; and the explanatory argument infers from failure of physical explanation to nonphysicality. One might say that these arguments infer from a failure of epistemic entailment to a failure of ontological entailment. The paradigmatic sort of ontological entailment is *necessitation*: P necessitates Q when the material conditional P⊃Q is metaphysically necessary, or when it is metaphysically impossible for P to hold without Q holding. It is widely agreed that materialism requires that P necessitates all truths (perhaps with minor qualifications). So if there are phenomenal truths Q that P does not necessitate, then materialism is false.

We might call these arguments *epistemic arguments* against materialism. Epistemic arguments arguably descend from Descartes's arguments against materialism (although these have a slightly different form), and are given their first thorough airing in Broad's book, which contains elements of all three arguments above.[9] The general form of an epistemic argument against materialism is as follows:

(1) There is an epistemic gap between physical and phenomenal truths.
(2) If there is an epistemic gap between physical and phenomenal truths, then there is an ontological gap, and materialism is false.

---

(3) Materialism is false.

Of course, this way of looking at things oversimplifies matters, and abstracts away from the differences between the arguments.[10] The same goes for the precise analysis in terms of implication and necessitation. Nevertheless, this analysis provides a useful lens through which to see what the arguments have in common, and through which to analyze various responses to the arguments.

There are roughly three ways that a materialist might resist the epistemic arguments. A type-A materialist denies that there is the relevant sort of epistemic gap. A type-B materialist accepts that there is an unclosable epistemic gap, but denies that there is an ontological gap. And a type-C materialist accepts that there is a deep epistemic gap, but holds that it will eventually be closed. In what follows, I discuss all three of these strategies.

5.4 Type-A Materialism

According to type-A materialism, there is no epistemic gap between physical and phenomenal truths; or at least, any apparent epistemic gap is easily closed. According to this view, it is not conceivable (at least on reflection) that there be duplicates of conscious beings that have absent or inverted conscious states. On this view, there are no phenomenal truths of which Mary is ignorant in principle from inside her black-and-white room (when she leaves the room, she gains at most an ability). And on this view, on reflection there is no "hard problem" of

explaining consciousness that remains once one has solved the easy problems of explaining the various cognitive, behavioral, and environmental functions.[11]

Type-A materialism sometimes takes the form of eliminativism, holding that consciousness does not exist, and that there are no phenomenal truths. It sometimes takes the form of analytic functionalism or logical behaviorism, holding that consciousness exists, where the concept of "consciousness" is defined in wholly functional or behavioral terms (e.g., where to be conscious might be to have certain sorts of access to information, and/or certain sorts of dispositions to make verbal reports). For our purposes, the difference between these two views can be seen as terminological. Both agree that we are conscious in the sense of having the functional capacities of access, report, control, and the like; and they agree that we are not conscious in any further (non-functionally defined) sense. The analytic functionalist thinks that ordinary terms such as "conscious" should be used in the first sort of sense (expressing a functional concept), while the eliminativist thinks that they should be used in the second. Beyond this terminological disagreement about the use of existing terms and concepts, the substance of the views is the same.

Some philosophers and scientists who do not explicitly embrace eliminativism, analytic functionalism, and the like are nevertheless recognizably type-A materialists. The characteristic feature of the type-A materialist is the view that on reflection there is nothing in the vicinity of consciousness that needs explaining over and above explaining the various functions: to explain these things is to explain everything in the vicinity that needs to be explained. The relevant functions may be quite subtle and complex, involving fine-grained capacities for access, self-monitoring, report, control, and their interaction, for example. They may also be taken to include all sorts of environmental relations. And the explanation of these functions will probably involve much neurobiological detail. So views that are put forward as rejecting functionalism on the grounds that it neglects biology or neglects the role of the environment may still be type-A views.

One might think that there is room in logical space for a view that denies even this sort of broadly functionalist view of consciousness, but still holds that there is no epistemic gap between physical and phenomenal truths. In practice, there appears to be little room for such a view, for reasons that I will discuss under type C, and there are few examples of such views in practice.[12] So I will take it for granted that a type-A view is one that holds that explaining the functions explains everything, and will class other views that hold that there is no unclosable epistemic gap under type C.

The obvious problem with type-A materialism is that it appears to deny the manifest. It is an uncontested truth that we have the various functional capacities of access, control, report, and the like, and these phenomena pose uncontested explananda (phenomena in need of explanation) for a science of consciousness. But in addition, it seems to be a further truth that we are conscious, and this phenomenon seems to pose a further explanandum. It is this explanandum that raises the interesting problems of consciousness. To flatly deny the further truth, or to deny without argument that there is a hard problem of consciousness over

and above the easy problems, would be to make a highly counterintuitive claim that begs the important questions. This is not to say that highly counterintuitive claims are always false, but they need to be supported by extremely strong arguments. So the crucial question is: are there any compelling *arguments* for the claim that, on reflection, explaining the functions explains everything?

Type-A materialists often argue by analogy. They point out that in other areas of science, we accept that explaining the various functions explains the phenomena, so we should accept the same here. In response, an opponent may well accept that in other domains the functions are all we need to explain. In explaining life, for example, the only phenomena that present themselves as needing explanation are phenomena of adaptation, growth, metabolism, reproduction, and so on, and there is nothing else that even calls out for explanation. But the opponent holds that the case of consciousness is different and possibly unique, precisely because there is something else, phenomenal experience, that calls out for explanation. The type-A materialist must either deny even the appearance of a further explanandum, which seems to deny the obvious, or accept the apparent disanalogy and give further substantial arguments for why, contrary to appearances, only the functions need to be explained.

At this point, type-A materialists often press a different sort of analogy, holding that at various points in the past, thinkers held that there was an analogous epistemic gap for other phenomena, but that these turned out to be physically explained. For example, Dennett (1996) suggests that a vitalist might have held that there was a further "hard problem" of life over and above explaining the biological function, but that this would have been misguided.

On examining the cases, however, the analogies do not support the type-A materialist. Vitalists typically *accepted*, implicitly or explicitly, that the biological functions in question were what needed explaining. Their vitalism arose because they thought that the functions (adaptation, growth, reproduction, and so on) would not be physically explained. So this is quite different from the case of consciousness. The disanalogy is very clear in the case of Broad. Broad was a vitalist about life, holding that the functions would require a non-mechanical explanation. But at the same time, he held that in the case of life, unlike the case of consciousness, the only evidence we have for the phenomenon is behavioral, and that "being alive" means exhibiting certain sorts of behavior. Other vitalists were less explicit, but very few of them held that something more than the functions needed explaining (except consciousness itself, in some cases). If a vitalist had held this, the obvious reply would have been that there is no reason to believe in such an explanandum. So there is no analogy here.[13]

So these arguments by analogy have no force for the type-A materialist. In other cases, it was always clear that structure and function exhausted the apparent explananda, apart from those tied directly to consciousness itself. So the type-A materialist needs to address the apparent further explanandum in the case of consciousness head on: either flatly denying it, or giving substantial arguments to dissolve it.

Some arguments for type-A materialists proceed indirectly, by pointing out the unsavory metaphysical or epistemological consequences of rejecting the view: e.g., that the rejection leads to dualism, or to problems involving knowledge of consciousness.[14] An opponent will either embrace the consequences or deny that they are consequences. As long as the consequences are not completely untenable, then for the type-A materialist to make progress, this sort of argument needs to be supplemented by a substantial direct argument against the further explanandum.

Such direct arguments are surprisingly hard to find. Many arguments for type-A materialism end up presupposing the conclusion at crucial points. For example, it is sometimes argued (e.g., Rey 1995) that there is no reason to postulate qualia, since they are not needed to explain behavior; but this argument presupposes that only behavior needs explaining. The opponent will hold that qualia are an explanandum in their own right. Similarly, Dennett's (1991) use of "heterophenomenology" (verbal reports) as the primary data to ground his theory of consciousness appears to rest on the assumption that these reports are what need explaining, or that the only "seemings" that need explaining are dispositions to react and report.

One way to argue for type-A materialism is to argue that there is some intermediate X such that (i) explaining functions suffices to explain X, and (ii) explaining X suffices to explain consciousness. One possible X here is *representation*: it is often held both that conscious states are representational states, representing things in the world, and that we can explain representation in functional terms. If so, it may seem to follow that we can explain consciousness in functional terms. On examination, though, this argument appeals to an ambiguity in the notion of representation. There is a notion of *functional representation*, on which P is represented roughly when a system responds to P and/or produces behavior appropriate for P. In this sense, explaining functioning may explain representation, but explaining representation does not explain consciousness. There is also a notion of *phenomenal representation*, on which P is represented roughly when a system has a conscious experience as if P. In this sense, explaining representation may explain consciousness, but explaining functioning does not explain representation. Either way, the epistemic gap between the functional and the phenomenal remains as wide as ever. Similar sorts of equivocation can be found with other X's that might be appealed to here, such as "perception" or "information."

Perhaps the most interesting arguments for type-A materialism are those that argue that we can give a physical explanation of our *beliefs* about consciousness, such as the belief that we are conscious, the belief that consciousness is a further explanandum, and the belief that consciousness is non-physical. From here it is argued that once we have explained the belief, we have done enough to explain, or to explain away, the phenomenon (e.g., Clark 2000, Dennett forthcoming). Here it is worth noting that this only works if the beliefs themselves are functionally analyzable; Chalmers (2002a) gives reason to deny this. But even if one accepts that beliefs are ultimately functional, this claim then reduces to the claim that explaining

our dispositions to talk about consciousness (and the like) explains everything. An opponent will deny this claim: explaining the dispositions to report may remove the third-person warrant (based on observation of others) for accepting a further explanandum, but it does not remove the crucial first-person warrant (from one's own case). Still, this is a strategy that deserves extended discussion.

At a certain point, the debate between type-A materialists and their opponents usually comes down to intuition: most centrally, the intuition that consciousness (in a non-functionally defined sense) exists, or that there is something that needs to be explained (over and above explaining the functions). This claim does not gain its support from argument, but from a sort of observation, along with rebuttal of counterarguments. The intuition appears to be shared by the large majority of philosophers, scientists, and others; and it is so strong that to deny it, a type-A materialist needs exceptionally powerful arguments. The result is that even among materialists, type-A materialists are a distinct minority.

5.5 Type-B Materialism[15]

According to type-B materialism, there is an epistemic gap between the physical and phenomenal domains, but there is no ontological gap. According to this view, zombies and the like are conceivable, but they are not metaphysically possible. On this view, Mary is ignorant of some phenomenal truths from inside her room, but nevertheless these truths concern an underlying physical reality (when she leaves the room, she learns old facts in a new way). And on this view, while there is a hard problem distinct from the easy problems, it does not correspond to a distinct ontological domain.

The most common form of type-B materialism holds that phenomenal states can be *identified* with certain physical or functional states. This identity is held to be analogous in certain respects (although perhaps not in all respects) with the identity between water and H_2O, or between genes and DNA.[16] These identities are not derived through conceptual analysis, but are discovered empirically: the concept *water* is different from the concept H_2O, but they are found to refer to the same thing in nature. On the type-B view, something similar applies to consciousness: the concept of consciousness is distinct from any physical or functional concepts, but we may discover empirically that these refer to the same thing in nature. In this way, we can explain why there is an epistemic gap between the physical and phenomenal domains, while denying any ontological gap. This yields the attractive possibility that we can acknowledge the deep epistemic problems of consciousness while retaining a materialist worldview.

Although such a view is attractive, it faces immediate difficulties. These difficulties stem from the fact that the character of the epistemic gap with consciousness seems to differ from that of epistemic gaps in other domains. For a start, there do not seem to be analogs of the epistemic arguments above in the cases of water,

genes, and so on. To explain genes, we merely have to explain why systems function a certain way in transmitting hereditary characteristics; to explain water, we have to explain why a substance has a certain objective structure and behavior. Given a complete physical description of the world, Mary would be able to deduce all the relevant truths about water and about genes, by deducing which systems have the appropriate structure and function. Finally, it seems that we cannot coherently conceive of a world physically identical to our own, in which there is no water, or in which there are no genes. So there is no epistemic gap between the *complete* physical truth about the world and the truth about water and genes that is analogous to the epistemic gap with consciousness.

(Except, perhaps, for epistemic gaps that derive from the epistemic gap for consciousness. For example, perhaps Mary could not deduce or explain the perceptual *appearance* of water from the physical truth about the world. But this would just be another instance of the problem we are concerned with, and so cannot help the type-B materialist.)

So it seems that there is something unique about the case of consciousness. We can put this by saying that while the identity between genes and DNA is empirical, it is not *epistemically primitive*: the identity is itself deducible from the complete physical truth about the world. By contrast, the type-B materialist must hold that the identification between consciousness and physical or functional states is epistemically primitive: the identity is not deducible from the complete physical truth. (If it were deducible, type-A materialism would be true instead.) So the identity between consciousness and a physical state will be a sort of primitive principle in one's theory of the world.

Here, one might suggest that something has gone wrong. Elsewhere, the only sort of place that one finds this sort of primitive principle is in the fundamental laws of physics. Indeed, it is often held that this sort of primitiveness – the inability to be deduced from more basic principles – is the mark of a fundamental law of nature. In effect, the type-B materialist recognizes a principle that has the epistemic status of a fundamental law, but gives it the ontological status of an identity. An opponent will hold that this move is more akin to theft than to honest toil: elsewhere, identifications are grounded in explanations, and primitive principles are acknowledged as fundamental laws.

It is natural to suggest that the same should apply here. If one acknowledges the epistemically primitive connection between physical states and consciousness as a fundamental law, it will follow that consciousness is distinct from any physical property, since fundamental laws always connect distinct properties. So the usual standard will lead to one of the non-reductive views discussed in the second half of this chapter. By contrast, the type-B materialist takes an observed connection between physical and phenomenal states, unexplainable in more basic terms, and suggests that it is an identity. This suggestion is made largely in order to preserve a prior commitment to materialism. Unless there is an independent case for primitive identities, the suggestion will seem at best ad hoc and mysterious, and at worst incoherent.

A type-B materialist might respond in various ways. First, some (e.g., Papineau 1993) suggest that identities do not *need* to be explained, so are always primitive. But we have seen that identities in other domains can at least be *deduced* from more basic truths, and so are not primitive in the relevant sense. Secondly, some (e.g., Block and Stalnaker 1999) suggest that even truths involving water and genes cannot be deduced from underlying physical truths. This matter is too complex to go into here (see Chalmers and Jackson 2001 for a response[17]), but one can note that the epistemic arguments outlined at the beginning suggest a very strong disanalogy between consciousness and other cases. Thirdly, some (e.g., Loar 1990/1997) acknowledge that identities involving consciousness are unlike other identities by being epistemically primitive, but seek to explain this uniqueness by appealing to unique features of the concept of consciousness. This response is perhaps the most interesting, and I will return to it.

There is another line that a type-B materialist can take. One can first note that an *identity* between consciousness and physical states is not strictly required for a materialist position. Rather, one can plausibly hold that materialism about consciousness simply requires that physical states *necessitate* phenomenal states, in that it is metaphysically impossible for the physical states to be present while the phenomenal states are absent or different. That is, materialism requires that entailments $P \supset Q$ be necessary, where P is the complete physical truth about the world and Q is an arbitrary phenomenal truth.

At this point, a type-B materialist can naturally appeal to the work of Kripke (1980), which suggests that some truths are necessarily true without being a priori. For example, Kripke suggests that "water is H_2O" is necessary – true in all possible worlds – but not knowable a priori. Here, a type-B materialist can suggest that $P \supset Q$ may be a Kripkean a posteriori necessity, like "water is H_2O" (though it should be noted that Kripke himself denies this claim). If so, then we would *expect* there to be an epistemic gap, since there is no a priori entailment from P to Q, but at the same time there will be no ontological gap. In this way, Kripke's work can seem to be just what the type-B materialist needs.

Here, some of the issues that arose previously arise again. One can argue that in other domains, necessities are not epistemically primitive. The necessary connection between water and H_2O may be a posteriori, but it can itself be deduced from a complete physical description of the world (one can deduce that water is identical to H_2O, from which it follows that water is necessarily H_2O). The same applies to the other necessities that Kripke discusses. By contrast, the type-B materialist must hold that the connection between physical states and consciousness is epistemically primitive, in that it cannot be deduced from the complete physical truth about the world. Again, one can suggest that this sort of primitive necessary connection is mysterious and ad hoc, and that the connection should instead be viewed as a fundamental law of nature.

I will discuss further problems with these necessities in the next section. But here, it is worth noting that there is a sense in which any type-B materialist position gives up on reductive explanation. Even if type-B materialism is true, we

cannot give consciousness the same sort of explanation that we give genes and the like, in purely physical terms. Rather, our explanation will always require explanatorily primitive principles to bridge the gap from the physical to the phenomenal. The *explanatory* structure of a theory of consciousness, on such a view, will be very much unlike that of a materialist theory in other domains, and very much like the explanatory structure of the non-reductive theories described below. By labeling these principles identities or necessities rather than laws, the view may preserve the letter of materialism; but by requiring primitive bridging principles, it sacrifices much of materialism's spirit.

5.6 The Two-Dimensional Argument Against Type-B Materialism

As discussed above, the type-B materialist holds that zombie worlds and the like are conceivable (there is no contradiction in $P\neg Q$) but are not metaphysically possible. That is, $P \supset Q$ is held to be an a posteriori necessity, akin to such a posteriori necessities as "water is H_2O." We can analyze this position in more depth by taking a closer look at the Kripkean cases of a posteriori necessity. This material is somewhat technical (hence the separate section) and can be skipped if necessary on a first reading.

It is often said that in Kripkean cases, conceivability does not entail possibility: it is conceivable that water is not H_2O (in that it is coherent to suppose that water is not H_2O), but it is not possible that water is not H_2O. But at the same time, it seems that there is *some* possibility in the vicinity of what one conceives. When one conceives that water is not H_2O, one conceives of a world W (the XYZ-world) in which the watery liquid in the oceans is not H_2O, but XYZ, say. There is no reason to doubt that the XYZ-world is metaphysically possible. If Kripke is correct, the XYZ-world is not correctly described as one in which water is XYZ. Nevertheless, this world is relevant to the truth of "water is XYZ" in a slightly different way, which can be brought out as follows.

One can say that the XYZ-world could *turn out* to be actual, in that for all we know a priori, the actual world is just like the XYZ-world. And one can say that *if* the XYZ-world turns out to be actual, it will turn out that water is XYZ. Similarly: if we hypothesize that the XYZ-world is actual, we should rationally conclude on that basis that water is not H_2O. That is, there is a deep *epistemic* connection between the XYZ-world and "water is not H_2O." Even Kripke allows that it is *epistemically possible* that water is not H_2O (in the broad sense that this is not ruled out a priori). It seems that the epistemic possibility that the XYZ-world is actual is a specific instance of the epistemic possibility that water is not H_2O.

Here, we adopt a special attitude to a world W. We think of W as an epistemic possibility: as a way the world might actually be. When we do this, we consider W *as actual*. When we think of W as actual, it may make a given sentence S true or