

Draft of January 25, 2005

## **Mental Maintenance: A Response to the Knowledge Argument**

**Jesse Prinz**

[jesse@subcortex.com](mailto:jesse@subcortex.com)

University of North Carolina at Chapel Hill

[Note to readers. This is a *very* rough draft of a chapter for a book manuscript in progress. My apologies for the umpteen typos, rough edges, terminological inconsistencies, missing references, and, especially, for any unintended misattributions. Please do not quote in print without consulting the final version. Also, note that the first two sections of the chapter cover a lot of familiar ground. Readers who are not interested in my misgivings about standard responses to the knowledge argument may want to skip to section 3.]

From an outside perspective, the philosophical literature on consciousness is often baffling. A disproportionate amount of ink has been dedicated to Frank Jackson's Knowledge Argument, and various related objections to physicalism, such as Nagel's bat argument, and a variety of arguments from conceivability and explanatory gaps. On cursory analysis, the Knowledge Argument is patently unsound. It illicitly draws a metaphysical conclusion from an epistemological premise. It doesn't take more than one philosophy class to learn that such arguments are fallacious. Why have philosophers wasted so much energy on a howler? The answer, I suspect, is that the obvious (and often less than obvious) responses to the Knowledge Argument don't succeed. The argument is remarkably resilient, and addressing it is a very useful way to draw positive conclusions about the nature of consciousness. When philosophers advance responses that seem to undermine the argument, new puzzles for physicalism almost immediately arise. Responding to Jackson is like killing a worm with a big rock: it's easy enough to kill the initial worm, but when you lift the rock, you expose a writhing tangle of new worms. That's what makes the Knowledge Argument so good.

In this chapter, I will argue that prevailing responses to the knowledge argument do not succeed. In discussing each solution, I will raise worries about the corresponding physicalist theory. I will also use the failings of each solution to identify a set of desiderata on an adequate materialist reply to Jackson. Then, with these desiderata in hand, I will propose a more successful solution that extends the AIR theory of consciousness. The proposal takes the form of theory of phenomenal knowledge (knowing what it's like). That theory is inspired by recent work in cognitive neuroscience, but it also incorporates positive lessons from the philosophical responses to Jackson that I criticize earlier in the chapter.

### **1. The Knowledge Argument**

In its simplest form, the knowledge argument begins with the premise that a person can know all the physical and functional facts correlated with a particular kind of phenomenal

experience without knowing what it is like to have that experience. From this, it is inferred that the experience in question is neither physical nor functional. If experiences are neither physical nor functional, then physicalism is false. Broad and Feigl devised thought experiments along these lines decades ago (Nida-Rümelin, 2002). Broad says that an archangel with microscopic vision would know the chemical structure of ammonia, but not its smell. Feigl says that a Martian might attain perfect knowledge of human behavior without knowing the qualitative character of human senses or sentiments. Nagel is responsible for bringing these kinds of considerations into the center of contemporary debates about consciousness. In “What Is It Like to Be a Bat?” he says that we can learn all about the physiology of echolocation without knowing “what it’s like” for a bat to echolocate (see also Farrell, 1950). Jackson added further refinements with two other cases. First, imagine Fred. He has four retinal color receptors rather than the usual three, and, as a result, he sees two distinct primary colors in the spectral range that we label red. We can’t see these two reds, and we cannot know what they are like by simply learning about the physiology of Fred’s visual system. Jackson’s second case concerns a brilliant neuroscientist named Mary. She has become a celebrity in the consciousness literature, and her story is tediously familiar.

Mary has lived her entire life in a black and white room with no access to colors. Nevertheless, she learns everything there is to know about what happens physically during color perception. Filling out some of the details, we might imagine that Mary first learns about what causes color experience. Light is electromagnetic radiation within a range that is detectable by human eyes. Wavelengths vary in size and the perceived color of an opaque object is determined by the wavelengths that it reflects, rather than absorbs, when illuminated by white light. Transparent objects transmit light, and luminous objects emit light. The ordinary human retina has three kinds of cone cells that maximally sensitive to three different wavelengths of light. The ratio of reflected light in these wavelengths is the primary determinant of perceived color. Retinal cells are stimulated in proportion to those ratios and then send a signal via ganglion cells in the optic nerve to the dorsal lateral geniculate nucleus (dLGN) of the thalamus. Some ganglion cells (the Magno cells) are fast and have low resolution; these are responsive to activity in retinal rod cells, which register light and dark, but not color. The Parvo cells are color sensitive, slow, and have high resolution. Magnocellular and parvocellular inputs arrive at different layers of the dLGN, which then feed forward to different layers of primary visual cortex (V1); color sensitive Parvo cells are localized to layers 2 and 3. The functional separation continues into V2, and then the parvocellular pathway feeds forward into various subregions of an area termed V4, in the extrastriate cortex. V4 is a likely locus of color experience (Zeki, \*\*\*). There is a subregion of V4 on the middle third of the lingual gyrus. Cortical color blindness (achromatopsia) occurs when this region is damaged (Heilman & Rothi, 1993). Mary knows all this. She also learns about the conditions that trigger cellular responses in V4, and she can plot a similarity space for those cellular responses. Mary also learns that V4 sends afferent signals to higher visual areas and areas associated with attention, orienting, and working memory in temporal, parietal, and frontal cortex. Convinced of the AIR theory of consciousness, she will infer that cellular response in V4 becomes conscious when activity levels reach the thresholds needed to send signals to working memory areas. To supplement her knowledge of the physiology, Mary also learns some psychological facts. For example, she learns that

seeing red causes an increase in heart rate, breathing, and muscle tension. She masters color vocabulary (which terms go together), and the colors of familiar objects.

In short, Mary learns everything physical about seeing red. No physically described fact escapes her. Nevertheless, Mary seems to be missing something. She does not know what it's like to see red (hereafter "WIL"). Upon her release from the room, she learns WIL; she sees roses, ripe tomatoes, strawberries, fire engines, splatter films, a torero's muleta, a Maine lobsters, and a fetching glass of Chatenuief de Pape. She didn't know what these things were like before her release, and she didn't know what it was like for others to see the mysterious color they call "red." Now Mary knows what red is like; "Red like *that*," she says while looking up at a cardinal in the park. Since Mary didn't know this before her release when she had merely physical knowledge of colors, WIL cannot be a physical fact. Phenomenal facts are evidently not physical, and, Jackson concludes, physicalism is false.

Mary has been much more popular than Jackson's Fred case. Perhaps readers have found Fred too fanciful. In actuality, the Mary case is a less realistic for two reasons. First, color cells respond to black and white stimuli, so Mary would be able imagine color after exposure to a black and white environment (even assuming she never caught a glimpse of her tongue or inner eye-lids). Second, color sensitive cells may need to be triggered during a critical period of development to become active. If Mary was really prevented from seeing light in the red range, she might never develop the capacity. Despite these difficulties with the case, Mary has certain advantages over Fred and the bats. Mary *can* ultimately overcome her ignorance of WIL without undergoing any unusual surgical procedure (sonar implants?!). Thus, she provides a useful way of posing the question of what's needed to transition from phenomenal ignorance into the glory of a Technicolor world. I will address that question below. But first, I want to summarize the argument and review standard strategies of response.

Here is a concise rendition of Jackson's argument:

- P1. Before her release, Mary knows all the physical facts about seeing red.
  - P2. Before her release, Mary does not know everything about seeing red. (She doesn't know what it's like to see red: WIL)
  - P3. If knowledge of WIL is not derivable from physical facts, then WIL is not physical.
- C. WIL is not physical, thus physicalism is false

The knowledge argument is irresistible to physicalists, because it is a bold attack that seems to rest a simple mistake. But *which* mistake? Critics of the knowledge argument agree that it is fatally flawed, but they disagree about where that flaw lies.

One response strategy that strikes me as unpromising is the rejection of premise 1 (though see Van Gulick, \*\*\*, for an heroic attempt). Ex hypothesi, Mary knows everything physical about seeing red. She could give a complete physical description of what goes on in the brain when red is perceived as well as the more abstract causal roles that the brain states in question play. The only room for dispute here concerns an

ambiguity in “physical fact.” Facts might be individuated by both sense (i.e., psychological concepts by which they are grasped) and reference (i.e., states of the world) or by reference alone (see Horgan, \*\*\*). If individuation requires modes of presentation, and there are physical states of the world that can be grasped via phenomenological states, then Mary does not know all the physical facts. This line of objection is unpromising, however, because the dualist can simply stipulate how the term “fact” is to be understood in the argument. Let a fact be a state of affairs, and let a physical fact be any state of affairs that can be described in the language of the completed physical science, using no phenomenological vocabulary (the assumption made by standard physicalists is that phenomenological properties will ultimately be reduced to some other properties). Let’s presume that Mary knows all *these* facts. P1 is secure. Most critics have focused on the other two premises, and I turn to these now.

## 2. Wrestling Mary

### 2.1 Imagination

Let’s begin with P2, which asserts that Mary does not know what it’s like to see red before her release. That seems extremely reasonable. Learning about the brain does not look like a sufficient way to attain phenomenological knowledge. But P2 has its critics.

One objection is advanced by Churchland (\*\*\*). He speculates that Mary might actually be able to imagine red before her release; by mastering neuroscience Mary might learn to enter novel brain states by act of will. This is completely implausible, I think, and unhelpful to the materialist. It is implausible because there is no special link between physiological knowledge and physiological control. We cannot digest our food better by learning about digestion. We know from the real world case of Knut Nordby, a congenitally colorblind vision scientist, that encyclopedic knowledge of color vision does not suffice for imaging red (Nordby, 1996; Sacks, 1996). Even if Mary could imagine red, she could not imagine what it is like for bats to see or what it is like for Fred, who has an extra rod cell. To know what the world is like for them, we would need new neural machinery, not willful control over the neural machinery already in place. Thus, Churchland’s reply has no hope of addressing the knowledge argument in its other guises.

In any case, Churchland’s suggestion would not help the physicalist (see Jackson, \*\*\*). The crucial claim in the knowledge argument is that knowledge of physical facts does not *constitute* complete knowledge of what it’s like. If that is the case, there is *prima facie* reason to think that what it’s like is not physical. Churchland’s imagination argument does nothing to block this inference. He merely says that knowledge of physical facts might allow us to arrive at knowledge of what it’s like through imagination. But this is consistent with the claim that the products of imagination are not physical. The physicalist must show that physical knowledge is, at least, knowledge of phenomenal qualities, not that it can be used to have phenomenal experiences. So Churchland’s imagination argument won’t suffice. In fairness, Churchland offers an alternative argument against Jackson, based on modes of presentation. I will raise worries about this strategy below.

In sum, it will not help to argue that Mary can imagine what red is like after learning everything about the brain. Indeed, turning Churchland’s suggestion on its head,

any adequate solution to the knowledge argument should explain why imaginative powers are so limited (as illustrated by the Fred case and the bat case). Usually knowledge facilitates imagination. If I tell you in detail about an exotic fish that lives in the Amazon, you'd probably be able to imagine it, but if I tell you about the electroreceptive sense organs of that fish you won't be able to imagine what it's like to be that fish. This gives us a desideratum:

**The Imagination Condition:** An adequate account of knowing WIL should explain why physical descriptions do not always allow us to imagine WIL.

We will encounter more adequacy conditions as we consider more unsuccessful responses to the knowledge argument. The next response has recently been advocated by Jackson himself.

## *2.2 Representationalism*

In the years since introducing the philosophical world to Mary, Jackson has had a change of heart. He now thinks that physicalism is defensible, but, for reasons we will examine below, he does not want to give up on P3 of the knowledge argument. Jackson thinks the best hope for defending physicalism is to show that Mary's knowledge of physical facts is sufficient for knowing WIL. Unlike Churchland, Jackson does not think that physical facts merely allow Mary to imagine red; he conjectures that there are physical facts the understanding of which constitutes knowledge of WIL.

Jackson's conjecture rests on two assumptions. The first is a thesis we have encountered in earlier chapters: representationalism. Representationalists claim that phenomenal facts are representational facts. More precisely, they think that every phenomenal quality of a mental state is constituted by representational properties of that state. Any two states that differ in their phenomenal qualities differ in virtue of representing different things. If phenomenal character can be understood in representational terms, then, says Jackson, there is hope for Mary. Jackson's second assumption is naturalism about representation. According to this thesis, representational properties are reducible to properties that can be specified in physical language. For example, representation may be explicable in terms of causal relations. On one popular naturalist story, a mental state represents whatever feature of the world caused it to be acquired. Another popular naturalist story says that mental states represent whatever reliably causes them to be tokened. There are numerous combinations, variants, and elaborations on these two ideas. Since causation in a physical process, causal theories of representation are physically specifiable.

Putting these two assumptions together, Jackson says that Mary is able to use physical knowledge to learn about all the properties that our phenomenal states represent. Mary can learn what red experiences represent and how they come to represent those experiences. Mary might learn, for example, that the experiences we call "red" are reliably tokened by encounters with certain reflectance properties. A visual experience of a ripe strawberry represents its surface as reflecting a preponderance of wavelengths in the range of 600 to 700 nanometers. Mary also learns that color experiences have a

variety of further representational and functional properties distinctive to sensory representations. Jackson thinks five properties are especially important:

1. Color experiences are rich (color is represented along with extension, location, motion, orientation, and shape);
2. That richness is inextricable (you cannot prize the color apart from the shape, etc.);
3. The experience is immediate (we do not perceive colors by first experiencing something else);
4. There is a “causal element” in the experience (we represent colors as located where we see them); and
5. Color experiences, like all perceptions, have a distinctive functional role (we use them to update our current beliefs).

Jackson supposes that this knowledge will be enough for Mary to know all the phenomenal facts about seeing red. The phenomenal qualities of red are exhausted by these representational and functional properties (for a similar claim, see Dretske, 1995). In presenting this account, Jackson admits that Mary will learn *something* on her release. He says she will acquire an ability to recognize, remember, and imagine colors (I will come to this proposal below). But he thinks those abilities do not constitute knowledge of how colors feel. The representationalist story is supposed to explain that. But does it?

I think Jackson’s proposal doesn’t pan out. First, his specific account of what sensory experiences are like is inadequate, and, second, representationalism is not a promising account of sensory qualities. I will, of necessity be, brief.

Jackson’s assumption that colors have representational content is very contentious. Color antirealists deny this, and secondary quality theorists argue that colors must be explained with reference to color experiences, not the other way around. But suppose we grant that color experiences represent. Could their representational content really explain how they feel? Obviously, the representational content is not sufficient. A verbal description of the reflectance properties of that red experiences register does not feel anything like a red experience, even though they represent the same property on Jackson account. To explain the feel of red, Jackson must appeal to the five properties just enumerated. The problem is that these five properties are neither necessary nor sufficient for color experience.

Let’s begin with Jackson’s claims that color experiences are rich, and inextricably so. This is often the case, but it’s hardly necessary. When we stare at a large uniform color field, we see no shapes orientation or motion, and when groups of colored shapes are presented very briefly, we can see the colors, but we are not sure which shapes they belong to (Treisman et al., \*\*\*). Consider also a patient named P.B. who sustained severe damage to his visual system as a result of an electric shock (Zeki et al., 1999). P.B. is essentially blind; he cannot discern shapes or locate features in space. But he can perceive and recognize colors. Clearly those experiences are not rich. P.B. also suggests that color experience need not be causal in Jackson’s sense. P.B. probably can’t locate colors at a causal source of their origin. Neither can we locate colors when, for example, we produce a red experience by pressing down on our eyes. Colors induced by psychoactive drugs may also be free-floating in this way. Must color experiences be

immediate? Probably not. Synesthetes experience colors when they see shapes, and ordinary perceivers can train themselves to imagine colors by association. Kosslyn et al. (2000) hypnotized subjects to experience colors when they looked at black and white images, and that resulted in activation of color centers in the brain. Presumably, these subjects were not seeing the colors immediately, but they were experiencing them. The final item of Jackson's list is functional role. Experiences, he says, have a characteristic role in belief updating. This strikes me as implausible. Experience can play many different roles with respect to belief; experiences can confirm, disconfirm, or be entirely ignored. There is no effect on belief that necessarily follows a red experience. Without spelling out the details, the functional role criterion cannot do any explanatory work.

I conclude that Jackson's five features are not necessary for color experiences. They are also insufficient. One can have a rich, immediate, causally-located, belief-influencing experience without any phenomenal qualities. For example, Haywood et al. (1994) describe a patient M.S. who can respond to color information, despite profound cortical achromatopsia (see Akins, \*\*\*, for a lengthy discussion). M.S. cannot recognize or, as far as we can tell, experience hues, but he can recognize and trace out the contour of shapes whose only difference from their backgrounds are chromatic. In other words, M.S. perceives color, and the color information is presumably immediate, rich, causally located, and capable of supporting spontaneously beliefs—but he has no color qualia!

The insufficiency of Jackson's five features can also be brought out in normal subjects. Consider unconscious priming. It seems plausible that we can have perceptual episodes that are rich, immediate, and so on, without awareness, when stimuli are rapidly presented and followed by a contrasting mask. There are even experimental demonstrations of unconscious color priming (Breitmeyer et al., 2004). Stretching things just a little bit, we can imagine a variant on the Mary case, called "Subliminal Mary." Before her release, Subliminal Mary is exposed to color patches, but only in a masked priming paradigm. Presumably she forms rich color representations during this procedure, but she still doesn't know what colors are like. Jackson's claim that Mary knows everything about colors prior to release seems grossly implausible.

These kinds of considerations bring out the limitations of representationalism. That theory is, at best, an account of how we ordinarily individuate phenomenal experiences, not an account of what phenomenal qualities are. Put in terms that I used in chapter 3, representationalism can help provide a theory of What we are conscious of, but not a theory of How we become conscious. But it is a terrible mistake to think that a theory of What we are conscious of is a theory of WIL. The qualitative character of experience is not exhausted by the representational content. Contents get their character by becoming conscious. What it's like to see red is a matter of red representations becoming conscious, not a matter of red representations as such. The case of Subliminal Mary shows this.

More realistic examples demonstrate that representational contents are not sufficient for distinguishing phenomenal qualities. Qualitatively different sensory experiences can have the same content. Consider sweetness. Gustatory sensations of sweetness are most reliably caused by sugars, which are chemical compounds consisting of chains of carbon, hydrogen, and oxygen at a ratio of 1:2:1. But the very same compounds can cause olfactory sensations of sweetness. They can even cause tactile sensations on the tongue, as when we assess the sugars in a wine by swishing it around

and sensing the mouth-feel. Thus, taste, smell, and touch are all representing the same property (the presence of CHO compounds), but they feel entirely different.

One might try to address these problems by looking for subtle differences in representational content. One suggestion is that taste, smell, and touch represent their respective input systems, in addition to representing features of the world; the taste of sugar would, on this picture, represent a CHO compound and the fact that the information is in a pathway that extends from the tongue. This move is hopelessly *ad hoc*. It's unlikely that, on any independently motivated theory of representation, the states of our sensory systems *represent* the systems of which they are states. This is no more plausible than the claim that the word "cow" represents both cows and the English language.

In any case, there is a principled reason for thinking that any attempt to individuate phenomenal quantities by appeal to their content is profoundly insufficient. Even if all phenomenal qualities could be distinguished by their representational content *as a matter of fact*, this must be a contingent fact. Most naturalistic theories of representation (what else can the physicalist appeal to?) are causal; they presume that reference works by standing in a causal relation to the external world. But, causal relations are always defeasible. We could, for example, take a brain in a vat, and get the visual states to register a random range of stimuli at the end of an electrode. We could wire the state that usually corresponds to red and the state that usually corresponds to green to exactly the same stimulus. We could also make the causes of neural events so random, that those events would end up representing nothing at all. A representationalist might insist that the brain events in question would not have the same phenomenal qualities as our brain events (because we use content to individuate brain states), but it would be desperate in the extreme to insist that such brain events had no qualities. And once we admit that those brain events have qualities, we are forced to admit that qualitative differences between them cannot be explained in representational terms.

There is only one retreat for the representationalist at this stage. She must drop the usual causal theories of reference and adopt some kind of internalism. Perhaps the representational content of an experience depends not on its external causes, but supervenes instead on things entirely in the head. This move is unpromising. First, since all our best independently motivated theories of representation are externalist, the shift to internalism is *ad hoc*. Second, representation is supposed to be a mind-world relation, and it is entirely unclear how any internal events could single out a unique content in the mind-external world. There are two kinds of internal properties that might be used to explain how phenomenal states represent. First, there are internal causal roles. The problem here is that internal causal roles are isomorphic with an open-ended range of properties in the world out there. Second, there are the phenomenal qualities themselves. Perhaps these are intrinsically representational (see, e.g., Loar, \*\*\*; Horgan and Tienson, \*\*\*; Searle, \*\*\*). I am very skeptical of this claim, because I think the apparent directedness of phenomenal states is an illusion brought on by their motor-affordances; we place visual qualities in the world, for example, because we are disposed to reach and orient towards them in particular ways. But, more to the point, if we must explain the representational content of sensory experiences by appeal to their phenomenal qualities, representationalism will be refuted; representationalism hopes the order of explanation will go the other way.

This critique of representationalism is too compressed to be decisive, but it should

raise the specter of doubt. Jackson's attempt to rescue physicalism from his own knowledge argument rests on a deeply problematic account of phenomenal qualities. We do, as a matter of fact, often use representational content to distinguish qualities, but that method of individuation is both contingent and insufficient. Representational content can vary, and there is more to qualitative character than content. What's more, representational content is not necessary for distinguishing qualitative states. A brain in a vat could make discrimination judgments even if its experiences had no content. This suggests that we have a method of distinguishing phenomenal states that does not depend on representational content. Most plausibly, we can distinguish these states by the vehicles that bear the content, not by the content itself. In other words, we distinguish phenomenal qualities by the representations that constitute them, not by what those representations happen to represent. In a slogan: we distinguish phenomenal qualities by representations, not representation.

This lesson echoes a claim that Fodor (\*\*\*) made in his early defenses in methodological solipsism. Psychological laws, he used to say, must be defined over internal states, so we must be able to individuate those states without reference to their external content. Applied to psychology *in general*, this precept is almost certainly mistaken (see Fodor, 1994), but it strikes me as plausible in the case of consciousness. I would advocate a corresponding desideratum on a solution to the knowledge argument:

**The Solipsism Condition:** On any adequate account, knowing WIL cannot (merely) be a matter of knowing what external features of mental states represent.

### 2.3 Abilities

I have been focusing on premise 2 of Jackson's knowledge argument. Most commentators agree that the offending premise is P3. According to that premise, WIL must be derivable from physical knowledge if WIL is itself physical. There are several popular strategies for rejecting this premise. I will consider these in turn.

The first strategy is to argue that knowledge of WIL is not factual knowledge; it is not knowing *that* something is the case. Instead it is procedural knowledge, or knowledge *how* to do certain things (Lewis, \*\*\*; Nemirow, \*\*\*). Know-how cannot be derived from knowledge that. We cannot learn how to ride a bike by reading about how legs move when riding; we must train our legs to move properly. Physicalism is fully compatible with the supposition that know-how requires experience. Like bike riding, Lewis and Nemirow suppose that knowing WIL is a form of know-how. In particular, knowing what red is like constituted by the ability to imagine red, recall red, and recognize red.

This proposal is deeply implausible. First, the kind of know-how in question is not necessary for knowing WIL. Some people are very bad at forming mental images. For them, imagining colors may (we can suppose) be impossible. Yet they clearly know what it's like to see red. Similarly, a person may be incapable of encoding or retrieving memories of experiences (consider anterograde amnesia), yet they can know what an experience is like while they are having it. Recognition is also unnecessary for knowing WIL. It is well known that discrimination outstrips recognition. We can experience

differences between two color patches, even when we cannot recognize, on a subsequent presentation of three color patches, which two patches we saw a moment earlier. Hasley and Chapanis (1951) estimated that we may be able to discriminate between one million different colors, but we can recognize only between eleven and sixteen!

The abilities in question are also insufficient for explaining knowledge of WIL. A computer program could be trained to recognize and recall red. That is, the program could be designed to store an internal record of spectral properties and call up that record when no spectral inputs were present. Yet it is easy to imagine a system doing this without having any qualitative experiences or subjective inner life. Defenders of the ability hypothesis would balk at this suggestion. The computer is not imagining *red* they will say, because *red* is a qualitative experience. But this reply is self-defeating. If the qualitative experience of red is not constituted by the abilities in question, then it is constituted by something else. If the qualitative experience is constituted by something else, then knowing what that experience is like can be explained by appealing to that something else (whatever it happens to be), and abilities will drop out of the picture.

This leaves us with a third desideratum:

**The Disability Condition:** On an adequate account, knowing WIL cannot require the capacity for imagination, recognition, or recall.

#### 2.4 Modes of Presentation I: Properties

There is a second strategy for blocking P3, which is by far the most popular response to the knowledge argument. Many of Jackson's critics think he has made a simple blunder. We cannot derive knowledge of WIL from knowledge of physical facts, because WIL is grasped using modes of presentation that differ from the kind used in our physical science (see, e.g., Carruthers, \*\*\*; Churchland, \*\*\*; Hill, \*\*\*; Loar, \*\*\*; Lycan, \*\*\*; Papineau, \*\*\*; Sturgeon, \*\*\*; Tye, \*\*\*). As Frege showed a century ago, whenever we have two modes of presentation for the same thing, we can have informative identities. These representational differences between physical concepts and phenomenal concepts block the inference from facts stated in brainy vocabulary to facts understood by means of phenomenal representations. These are the very same facts, but they are represented in different ways. This is a common occurrence. We cannot derive the identity between water and H<sub>2</sub>O because "water" and "H<sub>2</sub>O" are grasped using different concepts. Kripke (\*\*\*) added to this story by pointing out that we cannot always verify informative identities by doing *a priori* deductions, no matter how elaborate. Some identities must be verified *a posteriori*. In assuming that WIL should be derivable from physical facts, if WIL is physical, Jackson is assuming that theoretical identities must be discoverable *a priori*. If we simply suppose that phenomenal concepts and neural concepts refer via different modes of presentation, we can explain why Mary's physical knowledge does not tell her what it's like to see red.

How could Jackson make such an elementary mistake? The answer is that this seemingly obvious response the knowledge argument is deeply problematic on scrutiny. The appeal to modes of presentation faces serious objections. It will take some time to bring these out.

The troubles begin with a long-standing challenge to physicalism, which was influentially discussed by Smart (1959: n. 13), who attributed the objection to Max Black. The objection trades on a basic assumption about modes of presentation. Decedents on the Fregean tradition often assume that modes of presentation are individuated by properties; a mode presents its referent by representing some property of its referent. Thus, two co-referential modes differ if they present via different properties, and they are otherwise the same. From this it follows that, if there are two modes of representing the brain states that underlie conscious experience, they must represent those brain states via different properties. Neural descriptions quite obvious represent via neural properties. If, in addition to neural descriptions, we have a phenomenal way of representing phenomenal properties, then they must represent by something other than those neural properties. So there must be phenomenal properties in addition to neural properties. Even if experiences are neural states, those neural states must have nonphysical properties. Property identity theory looms.

Smart's own reply is to argue that phenomenal modes of presentation represent in a topic neutral way. He admits that they do not represent phenomenal states by explicitly describing their neural properties, but neither do they make explicit appeal to any properties that are mental. Rather, they present experiences to us by appeal to the role those experiences play: red is the kind of experience I have when I see strawberries or ripe tomatoes under good lighting conditions.

Smart's strategy faces an obvious objection. The topic-neutral properties that he mentions are functional role properties. His suggestion is that we pick out our phenomenal states by the role they play. But this seems incredibly unlikely, at least if we consider the kinds of roles that Smart invokes. A person can know what red experiences are like without having any idea about how those experiences are normally elicited. Recall, for example, the neurological patient who sees color but no form. A person born with this condition might have no way of identifying the kinds of things that are red. Red experiences may play a more subtle or sophisticated functional role. Perhaps the role involves contributions to state transitions in our visual systems that would have to be identified by scientific psychology. But notice two things. First, these fancy functional roles are far from obvious, yet it is perfectly obvious what red is like. Second, Mary could have a thorough description of those roles and still not know what red is like. This suggests that our mode of presentation for red is not merely a role concept.

Once we rule out the role strategy, Max Black's objection seems to re-appear. If the property by which we pick out phenomenal states is not structural (i.e., neural) or functional, then it must be irreducibly phenomenal. We are back with property dualism. Kripke famously strengthens this argument by pointing out that phenomenal modes of presentation and neural modes of presentation both represent via essential properties (properties that are necessary and sufficient for identity across all possible worlds). So if phenomenal and physical modes of presentation are different, then there can be no true psychophysical identities. We are stuck, not with property dualism, but with full on dualism about qualitative mental states. I won't be much concerned with the modal move here. Even without it, friends of the mode or presentation view face a formidable hurdle. If modes are individuated by properties, then it would appear that physicalists cannot invoke phenomenal modes without undermining their cause.

Faced with this objection, friends of modes of presentation can respond in one of

two ways. Either, they can argue that phenomenal modes of presentation present via properties that can be accommodated within physicalism, or they can argue that we ought not individuate modes of presentation by properties. I will discuss a version of this second strategy for the next section, and I focus on the accommodation move here.

The accommodation strategy for tackling the knowledge argument goes like this. Admittedly Mary acquires a new mode of presentation when she leaves the room, and admittedly it refers via a property. Moreover, that property is a different property than the one she was able to represent using a neural description, but different only in an innocuous sense. Physicalism allows for levels of analysis. With different levels, there can be a proliferation of properties. There are properties of physics, properties of chemistry, properties of biology, and properties of various social sciences. This proliferation is not an affront to physicalism, however, because the properties at each level can be accommodated by lower levels. By “accommodation,” I mean to refer to a variety of different relationships between higher- and lower-levels. In some cases higher-level properties are reducible to lower-level properties, in other cases, they are constituted by lower-level properties, and in still other cases they are realized by lower-level properties. In each case, physicalism remains untarnished. We can characterize physicalism (at least approximately) as the view that a world with exactly the same facts at the level of physics would necessarily have the same facts at each level at which there are true facts in this world (barring complications of haecceity). High-level properties that are reduced, realized, or constituted by low-level properties are unproblematic because, when we fix low-level properties, they always come along for the ride.

There are different ways to think about what it means to say that higher-level properties are different from lower-level properties. In the case of constitution and realization, we might say that the higher-level properties enjoy some degree of autonomy. Their status as properties comes from the existence of generalizations that would be lost if we were to descend to lower-levels. In the case of reduction, higher-level properties are different in a pleonastic sense; they are differentiated by the fact that we pick them out using different vocabularies. Frugal realists might regard higher-level properties with some skepticism, dismissing them as hypostases in the pejorative sense. That need not bother us here. The present point is that there is at least a vulgar use of the term “property” that allows us to say that high-level properties exist and, most importantly, their existence is no problem for physicalism.

The upshot is this. The physicalist can respond to the Max Black objection by arguing that phenomenal modes of presentation are individuated by properties that can be physically accommodated. Smart’s topic neutral approach, which I dismissed, is a version of this strategy. Physicalists who do not buy into Smart’s proposal might nevertheless hold out hope for some other solution along these lines. Many physicalists assume that mental properties are designated by modes of presentation that refer via properties that can be physically accommodated. Call this the accommodation assumption. To make this work against the knowledge argument, physicalists need to defend one more assumption: they need to claim that some properties that can be physically accommodated *cannot be logically deduced* from descriptions of facts at lower levels of analysis. Call this the *a posteriori* physicalism assumption. *A posteriori* physicalism has been widely accepted since Kripke (\*\*\*) presented his account of theoretical identities. Water is H<sub>2</sub>O but, intuitively, we cannot deduce this fact *a priori*.

If properties that can be physically accommodated cannot be physically deduced, then we can explain Mary's predicament without violating the strictures of physicalism. Her neural knowledge cannot be used to derive knowledge of WIL, even though facts about WIL can be physically accommodated.

Jackson (\*\*\*) thinks that there is a principled reason for thinking this strategy will never work (see also Chalmers, \*\*\*; Jackson and Chalmers, \*\*\*). In particular, he rejects the *a posteriori* physicalism assumption. Of course, Jackson acknowledges that many identity statements are not true analytically. When there are two different modes of presentation for the same thing, we cannot always infer the identity by simply examining those modes of presentation and nothing else. But that does not mean the identity cannot be discovered *a priori* when we include other information. Return to the water case. Water has a mode of presentation that fixes reference by means of stereotypical properties. We conceive of water as a clear liquid that boils when heated. We know this stereotype *a priori*. Now suppose we supplement this *a priori* knowledge with complete knowledge of microphysics. The latter knowledge is sufficient to logically deduce that H<sub>2</sub>O satisfies the water stereotype: it is a liquid at room temperature, it transmits light rather than reflecting it, and bubbles upward when its kinetic energy rises. Thus, microphysics, plus knowledge of the water stereotype is enough to deduce that water is H<sub>2</sub>O, because we can deduce that H<sub>2</sub>O has the properties that are linked *a priori* to water.

In stark contrast, microphysics does not help Mary deduce the alleged identity between red experiences and neuronal activations in V4, because she cannot deduce that such neuronal activations satisfy the red stereotype. If red experiences can be said to have a stereotype (a set of features by which we recognize that we are having such an experience) those features must themselves be phenomenal. And, as we have already seen, there is no way to deduce that brain states have any phenomenal qualities. Thus, there is a disanalogy between the case of ordinary psychophysical identities, which can be deduced *a priori* from stereotypes and microphysics, and psychophysical identities, which cannot be deduced. If *a posteriori* physicalism is false, physicalism is in jeopardy. Here's the argument (associated with Jackson and Chalmers):

JC-P1. Facts represented via phenomenal concepts (phenomenal facts) cannot be deduced from facts represented via physical concepts (physical facts).

JC-P2. If phenomenal facts are physical, then they can be deduced from a complete set of facts represented via physical concepts. (Entailed by *a priori* physicalism.)

JC-C. Therefore, phenomenal facts are nonphysical.

One can reply to this latest epicycle in the knowledge argument by rejecting *a priori* physicalism. Block and Stalnaker (\*\*\*) offer a trenchant critique. They argue that the water case is more like the color case than Jackson claims: there is no way to deduce the identity between water and H<sub>2</sub>O from microphysics alone. To establish a disanalogy between the water case and the case of psychophysical identities, Jackson says we can prove that water=H<sub>2</sub>O *a priori*. His argument is based on two assumptions:

(JA1) It is an a priori truth that water is that which plays the role spelled out in the water stereotype; and

(JA2) Microphysical facts entail that H<sub>2</sub>O plays the water role.

Against (JA1), Block and Stalnaker argue that there is a more complicated relationship between concepts and stereotypes. Some concepts refer to the natural kind that fills the associated stereotype role, but, when there is no single natural kind that does that, then a concept may refer instead to the stereotype role itself, rather than the role fillers. If there had been both H<sub>2</sub>O and XYZ on Earth, we might have concluded that water is just any stuff that's clear, wet, and boils when heated. We do not know that water refers to the natural kind that plays the water role *a priori*, because it might not refer to any natural kind at all. Against (JA2), Block and Stalnaker argue that microphysics cannot be used to draw conclusions about substances stated in folk vocabulary. The water stereotype includes features like clear, wet, and boils. These are not terms from microphysics, so we cannot deduce them from microphysics. We can deduce that H<sub>2</sub>O molecules rush to the surface when their mean molecular kinetic energy increases, but we cannot deduce that this counts as boiling *a priori*. "Boiling" may be a natural kind term. Had we been in another world, it would have referred a different process. So we cannot know *a priori* that the behavior observed in H<sub>2</sub>O is an example of boiling. Instead, we notice that H<sub>2</sub>O molecules rush to the surface in just those cases where water boils, and we infer that H<sub>2</sub>O and water are identical; that is the simplest explanation. This inference is *a posteriori* and it is based on a correlation. But that is exactly how we come to identify mental states with brain states. The two co-occur, and we presume they are identical because that's the simplest explanation. So *a priori* physicalism is not true in the water case, and the water case and the psychophysical case are closely parallel.

This rendition cannot do justice to all the moves made by Block and Stalnaker in their densely argued paper, but it will suffice for present purposes. I agree with their conclusions about *a priori* physicalism on general grounds. Post-Quinean scruples weigh heavily against *a priori* analysis of concepts, even when those analyses are given in terms of stereotypes. Stereotypes are learned tools for rough and ready categorization, which are revisable, and possibly false. Moreover, figuring out what a concept designates is, in the face of massive indeterminacy, often a matter of decision, not discovery. Does "water" refer to isotopes? Does "tiger" refer by clade or by genotype? Are wolves dogs? Are platypuses mammals? Are viruses alive? Nothing in microphysics has any hope of answering these questions.

For all these reasons, I think that Jackson's attempt to vindicate Max Black's objection won't succeed. Unfortunately, however, the story doesn't end there. Even if *a priori* physicalism is false, there is still a related argument against the mode of presentation strategy, which is more difficult to circumvent.

Let's assume with Block and Stalnaker that there is no analytic connection between "water" and "filler of the water" role. Let's also assume that there is no deductive argument for the conclusion that H<sub>2</sub>O plays the water role, because that role may be specified using natural kind terms. But instead of the water role, let's talk about something very similar. Each concept comprising a stereotype of the water role, like BOILS and LIQUID, is associated with a set of macroproperties that we can, for lack of a

better term, call appearances. Appearances are not mental states; they are just the superficial properties by which we would identify something as boiling or being a liquid. At the microlevel appearances are multiply realized. Granular solids and liquids might both have a liquid appearance. To a first approximation, we can think of an appearance as how something behaves at a high-level of analysis. When we investigate microphysical entities, we can draw inferences about their appearances. This style of inference is not deductive, because microphysics may contain no appearance predicates, but it is a perfectly tractable ampliative inference procedure. For example, we might use a combination of mereological and mechanistic reasoning. We see micro-entities grouping together, and we can label those groups. Then we describe the behavior of groups of micro-entities, in much the way we might talk about the trajectory of a group of fish. We can describe interactions between groups, and groups of groups. We can use the same causal vocabulary that we use to describe local micro-interactions to describe macro-interactions. Basically, the idea is that we look for increasingly complex *Gestalten*. I will use the term “gestalting” to refer to this kind of inference procedure.

With gestalting, we can draw inferences about how micro-entities appear. Consider an example. When a collection of H<sub>2</sub>O molecules comes into contact with a collection of sparsely arranged molecules, and then moves away, many of the H<sub>2</sub>O molecules stay behind, intermingling with the other collection. Labeling this, we might say that H<sub>2</sub>O is absorbed\* by porous\* materials. “Absorbed\*” and “porous\*” are labels for appearance properties. They are fully gestaltable from an analysis of H<sub>2</sub>O. We also note that H<sub>2</sub>O molecules are far apart, and not strongly bonded at prevailing ambient temperatures. Call this “wetness\*.” In addition, H<sub>2</sub>O molecules transmit light, rather than merely reflecting it; so items placed behind a collection of H<sub>2</sub>O molecules reflect wavelengths through water. Let’s call this “transparency\*.” Finally, collections of H<sub>2</sub>O molecules rise to the surface when they come into contact with a something that increases their kinetic energy. Call this “boiling\*.” In this way, we can derive the fact that H<sub>2</sub>O is wet\*, transparent\*, and boils\*.

Of course, we don’t know that H<sub>2</sub>O is wet, transparent, and boils, because each of these features could be a natural kind, and there are possible worlds where “wetness,” “transparency,” and “boiling” denote different microproperties than those that I have described. Apparent wetness can be a feature of granular solids. Apparent transparency can be a feature of tiny mirrors that reflect light. Apparent boiling could be explained by an invisible pulley system that hoists small balls of water to the surface. Therefore, we cannot deduce that H<sub>2</sub>O plays the water role, as opposed to merely being a water look-alike. But, as we will see, this may not matter for the knowledge argument.

It seems reasonable to assume that every physical property is gestaltable. Physical things are just material entities, collections of material entities, or activities of material entities. It seems that all these things are gestalts of microphysical facts. When we gestalt, we may discover some “real patterns” that are (or could be) realized by different microphysical entities, and we may, therefore, adopt an ontology that allows for multiple realizable macroproperties. But real patterns are *Gestalten*. If we were to add properties to our ontology that were not *Gestalten*, we would end up with properties that could change even if we kept every material entity and activity fixed. Such properties would be nonphysical. There are two reasons why we cannot deduce that H<sub>2</sub>O is water. We cannot be sure that H<sub>2</sub>O plays the water role, as opposed to the water-appearance

role, and we cannot be sure that “water” refers to the stuff that plays the water role. But we can be sure that if “water” refers to physical stuff, that stuff can be identified by gestalting. Physicalism about water does not entail the *a priori* derivability of the conclusion that water=H<sub>2</sub>O, but it does entail that “water” refers to gestalttable stuff.

Likewise, physicalism entails that phenomenal experiences are gestalttable. This can be captured by the following gestaltability principle:

(GP) If WIL is physical, then WIL is gestalttable from physical facts.

Now (GP) looks like a principle that might be used against physicalism by denial of the consequent. If dualists assert that WIL is not gestalttable, they can refute physicalism. The trouble is that one can’t do that without begging the question. Physicalists think that WIL is a brain state, and brain states are certainly gestalttable. To avoid question-begging, dualists should start with an epistemic version of the gestaltability principle:

(GP') If WIL is physical, then knowledge of WIL is gestalttable from physical facts.

*Knowledge* of WIL certainly isn’t gestalttable. If it were, Mary would be able to learn what red is like, because she can gestalt to her heart’s content. So (GP') would be a very powerful tool against the physicalist. Indeed, (GP') is just a variant of (P3) in the knowledge argument. The contrapositive of (GP') says, “If knowledge of WIL is not gestalttable from physical facts, the WIL is not physical.” Compare this to:

P3. If knowledge of WIL is not derivable from physical facts, then WIL is not a physical.

If we interpret “derivable” as “deductively deducible *or gestalttable*,” then P3 and (GP') are nearly synonymous. If we could prove that (GP') is true, we could get the knowledge argument off the ground, because Mary cannot gestalt knowledge of WIL from physical facts. But how do we argue for (GP')? Here’s a plausible derivation:

(GA-P1) If a property is physical, then it is gestalttable from physical facts.

(GA-P2) Knowledge of WIL consists in having a mode of presentation that presents WIL via some property. (The standard physicalist mode of presentation reply to the knowledge argument.)

Therefore:

(GA-C) If WIL is physical, then knowledge of WIL is gestalttable from physical facts. (=GP')

Stated more informally, if my knowing WIL is having WIL presented to via a property, and all physical properties are gestalttable, then I should be able to discover WIL by gestalting if WIL is physical.

This has been a tangled journey, but the upshot is simple enough. On the standard mode of presentation proposal, physicalists claim that knowledge of WIL is mediated by modes of presentations of properties. But, if modes of presentation are mental graspings of properties, and those properties cannot be inferred by gestalting from physical facts, then those properties are not physical. Far from undermining the knowledge argument, the standard mode of presentation view invites it. Max Black's suggestion that modes of presentation proliferate properties has been vindicated.

To avoid this conclusion, one can argue that modes of presentation do not introduce properties (they are individuated in some other way). I will examine two versions of this option below, but first a desideratum:

**The No Property Condition:** An adequate account of knowing WIL cannot assume that Mary represents her experience of red via a property.

### *2.5 Modes of Presentation 2: Mentalese Syntax*

One strategy for saving modes of presentation is to give up on the Fregean axiom that every mode presents by means of a property. Perhaps there is some other way to individuate modes. Frege favored a cognitive significance test: two modes differ if they can flank an informative identity. But Frege's examples of co-referential modes vaguely suggest that he thought cognitive significance was explicable in terms of reference-determining properties. With due respect to Frege, this might not be the case. It seems that there could be concepts that differ in cognitive significance (a psychological fact), but are alike in what properties they denote and the in the properties by which they denote (a semantic fact). Indeed, two tokens of the same proper name can have different cognitive significance. After hearing the name Paderewski on two occasions, one might wonder whether there is one person with this name or two: "Is Paderewski Paderewski?" (Kripke, \*\*\*; see also Block, \*\*\*).

If we don't individuate modes of presentation by properties, then how do we individuate them? The question can also be put in terms of concepts. If concepts are not individuated by properties (other than their referents), then how are they individuated? One suggestion is advanced by Fodor (\*\*\*). He is a semantic atomist. For him, concepts expressed by a single word are not semantically related to other concepts; one can have a single concept and no others. To explain cognitive significance, he suggest that concepts have syntax. Just as public words have shapes or sounds ("Paderewski"), concepts are words in a language of thought that have "formal" properties. Fodor's proposal is that the formal properties of a concept are determined by its causal role in the mind. Each concept is a symbol in a symbol system, and each gets its syntactic identity, but not its meaning, from how it related causally to other symbols in the system. These relations do not determine reference, but they do determine cognitive significance. Thus, we should not say that co-referential concepts with different significance differ in the properties by which they refer.

Something like this idea has been invoked by Lycan (\*\*\*) to address issues surrounding the knowledge argument. He thinks the mind contains one symbol system associated with conscious experience, and another separate symbol system associated

with linguiform thoughts. Mental symbols get their cognitive significance from the roles they play within a system. This has two consequences. First, co-referential symbols across the two systems need not refer by presenting different properties. A phenomenal concept and a neural concept do not present different properties; they just have different syntax. This overcomes Max Black's worry about property proliferation. Second, these symbols are not intertranslatable; if a symbol in one system gets its meaning by its causal role in that system, and no symbol in the other system has a comparable causal role, then there will be no way to capture the significance of one symbol using the other. This, Lycan argues can help why Mary can't derive phenomenal knowledge from physical facts. There is no way to express, much less derive, one by means of the other, because they are represented in fundamentally different mental languages. Mary can know everything about the brain without being able to formulate any thoughts in her dormant color code. Lycan does not elaborate how these symbol systems work, but, taking some liberties, one might suggest that the color code contains symbols that get their significance by their causal effects on each other. The red symbol is the one that, when co-activated with the blue symbol, outputs the violet symbol, which is never spatially coincident with the yellow symbol, and so on.

I'm not sure if this does justice to Lycan's proposal, but if it does, the proposal faces some serious objections. First, the kinds of causal relations between colors that I mentioned can be modeled by an isomorphic system of rules between arbitrary symbols. Each color can be mapped onto a name. Mary can learn that "red" and "blue" cause "violet." Rather than merely representing these causal relations, she might internalize the transformation rules, so that thinking "red" and "blue" concurrently would cause her to think "violet." If cognitive significance were simply a matter of adopting a symbol system with the right causal relations, Mary's mastery of these transformation rules should suffice for learning what colors are like.

Second, it seems intuitively implausible that experiences have their cognitive significance in virtue of causal relations of the kind under consideration. An experience seems to have its significance in virtue of the phenomenal character it has, and that character can occur in the mind even when the causal relations in question are not being realized. Causal relations are dispositional, but phenomenal character can be experienced in a temporal instant.

Third, the idea that color representations have their cognitive significance in virtue of their relations to other color representations implies that two individuals with a different set of color representations will not have *any* color experiences in common, because no pair of color representations in these two systems will have the same role. There is empirical reason for doubting this. Graham and Hsia (1958) present data on a woman who is dichromatic (specifically, a deuteranope) in one eye and trichromatic in the other. They ask her to match color sample presented to each eye separately. The crucial finding is that she has no difficulty seeing sample presented in one eye a just like samples presented in the other. She makes many errors of course, but there are also samples that she matches correctly across. The main point is that visual experiences in a dichromatic system can be mapped onto experiences in trichromatic system, which is at odds with the hypothesis that we grasp color experiences using representations that are individuated by relation to other color representations.

In sum, it is unlikely that our representations of phenomenal states are

individuated by their causal relations to other phenomenal representations. That said, I would not rule out that phenomenal representations are individuated by some other aspect of their causal role. There may be some promise in this general strategy. But notice the power of the third objection to the color code proposal that I have just been considering. We need a causal role story that does not assume experiential holism. Experiential holism is the view that the character of any given experience depends on the character of other experiences of the same type. If experiential holism were true, then two people with a different range of color discrimination capacities would experience *each* color differently. The case of unilateral color blindness suggests that experiential holism is false. Indeed, it might be the case that a person could experience one visual quale and no others. Consider GY, the famous blindsight patient, who can experience a subtle sensation in his “blind” field when present with a moving high contrast gradient. By comparing this sensation to his intact field, Stoerig (\*\*\*) determined that it is a subtle sensation of motion. That is evidently *all* GY can experience in his “blind” field. In this respect, visual experiences seem to be atomic, not holistic.

This gives us another desideratum:

**The Independence Condition:** On an adequate theory, phenomenal concepts cannot be individuated by relations to other phenomenal concepts (i.e., holistically).

### *2.6 Modes of Presentation 3: Recognitional Concepts*

Some commentators on the knowledge argument think that we can defend physicalism by postulating a class of concepts that refer by means of direct causal connections to what they represent. A natural strategy is to look for an internal analogue of pointing. When we point to an object, we don’t need to describe it. In languages, we have a class of expressions that refer in this way called demonstratives, such as “this” and “that.” For a linguistic demonstrative to refer, the word must be accompanied by some way of drawing attention to the designated object. Pointing can play a role in this, but pointing does not always suffice. If I point to a cluttered table and say “that,” my gesture will not fix reference. Kaplan (\*\*\*) calls the required additional component a demonstration. Demonstrations often include mental images. If say “that” while forming an image of a book on the table, I will have succeeded in referring to the book, whereas I would have referred to the brandy snifter on the table, had I formed an image of it instead of the book. Suppose, however, that I refer demonstratively to my own inner experiences. Suppose I form a thought expressed by “red is like *this*,” while focusing on a red experience. Here, there is no need for a demonstration. The concept expressed by “this” can refer directly, because it is directly wired to the inner state in question. Inner demonstratives can refer without describing any features of their referents.

Brian Loar has developed a sophisticated suggestion along these lines (\*\*\*). Loar begins by exposing a semantic premise that is implicit in Jackson’s argument (and in Kripke’s modal argument). Jackson assumes that any seemingly contingent identity claim can be true only if one of the concepts used to grasp that claim refers by means of a contingent property. If this premise is accepted, physicalism is doomed, because

phenomenal concepts do not refer via contingent properties. Loar (1990) argues that the semantic premise is false.

To make this case, Loar introduces a class of recognitional concepts, which denote without describing. Recognitional concepts have no meaningful semantic analyses, nor any semantically interpretable mental states mediating between them and what they represent. Consequently, a recognitional concept cannot be deduced from any other concept that happens to refer to the same thing. It follows that identity conjectures containing recognitional concepts are *a posteriori*. But, unlike concepts in other non-obvious identity conjectures, recognitional concepts do not refer via contingent properties. They refer via direct causal links directly to the things that they are used to recognize. Thus recognitional concepts provide a counterexample to the semantic premise that all seemingly contingent identities involve concepts that refer via contingent properties.

To save physicalism, Loar simply proposes that we represent our phenomenal states by means of recognitional concepts. How do we know that we are having a red experience? We have a recognitional concept that is directly caused by that experience, and the recognitional concept does not describe the experience. Suppose now that red experiences are merely brain states. The recognitional concept used to detect such experiences would not reveal that fact to us, nor any other. Recognitional concepts do not describe what they denote. So when we entertain the question, “is *this* experience a brain state?” the answer is not obvious. The phenomenal concept that we use to pick out the experience is silent on that question. Thus, the non-obviousness of psychophysical identities is explained, and the explanation is completely consistent with physicalism. Loar also uses this strategy to address the knowledge argument. Jackson assumes that, if Mary cannot deduce knowledge of WIL from knowledge of brain states, then WIL is not a brain state. But, on Loar’s view, knowing WIL is a matter of having a capacity to recognize phenomenal states by means of recognitional concepts. Since recognitional concepts are non-descriptive, knowledge couched in terms of them cannot be deduced from knowledge couched in other concepts. Therefore, there is a principled reason for thinking that we should not be able to deduce knowledge of WIL from knowledge of brain states that we grasp using scientific concepts. This inferential barrier poses no threat to the physicalist.

Loar’s response has advantages over the other two modes of presentation views we have considered. Unlike the first approach, he does not assume that we represent phenomenal states by means of their properties, so he appears to be immune to the Max Black Objection. And, unlike the Mentalese syntax approach, he does not assume that individuation of phenomenal concepts is holistic.

Still, Loar’s response may be problematic. One line of objection has been put forward by Chalmers (\*\*\*). He finds a tension between the two major components of Loar’s account of phenomenal concepts. On the one hand, these concepts are supposed to be non-descriptive. They refer directly, not via any specific properties of the phenomenal states that they designate. Chalmers calls this neutrality. On the other hand, phenomenal concepts are supposed to be opaque: they cannot be freely replaced with co-referring concepts, even if those co-referring concepts refer via essential properties. Substitution with co-referring concepts changes cognitive significance. Mary can believe that “red experiences” are identical with V4 activations, without know that red

experiences are like *that*, where “that” expresses a phenomenal concept. Normally, differences in cognitive significance between concepts are explained by presuming that that those concepts refer via different properties. Opacity presupposes non-neutrality. Loar owes us an explanation, says Chalmers, of how a concept can be both neutral and opaque.

I think this objection can be answered. Opacity does not require non-neutrality. In principle, there can be an informative identity claim made using any two representations that are not token identical, as Paderewski cases illustrate. Opacity is cheap. A label that tells us nothing about its referent provides us with no information about which we can derive identities using other labels. Neutrality yields an intractable form of opacity. There is no tension here at all. Chalmers has not identified a fatal objection to Loar.

But Loar is not off the hook. His account seems to face the following dilemma. As I have been presenting Loar’s view, we are to imagine that phenomenal concepts are radically neutral; they are inner labels that tell us nothing about the phenomenal states that they designate. This is seriously implausible. When we say, “That’s red,” we are not reporting that a neutral label has switched on in our heads; we are reporting that we recognize a specific phenomenal quality—redness. If phenomenal concepts were neutral labels, then we couldn’t use them to draw substantive conclusions about our phenomenal states, such as the conclusion that orange is more like red than it is like blue. Moreover, if phenomenal concepts were neutral labels, then one could use them in the absence of the phenomenal states that they designate. We could think, “red is like *this*,” without imagining or seeing red. That is an unacceptable consequence. Reflecting on what red is like requires having a red experience. The experience itself is a constituent part of how we grasp it. The cognitive significance of a phenomenal concept of red experience must be explained, at least in part, by appeal to the experience itself. I think this is a desideratum on an adequate response to the knowledge argument:

**The Phenomenal Constituent Condition:** An account of knowing WIL to have an experience should assume that the knowledge is constituted in part by the experience itself.

I think that Loar may actually agree with this condition. His discussion sometimes gives that impression. If this is right, then Loar does not really accept the neutrality claim. Rather, he thinks that phenomenal concepts refer by presenting they very properties that they designate. We point to our red experiences by means of those very experiences. But if this is Loar’s view, then it is equivalent to the claim that phenomenal concepts are individuated via phenomenal properties. On this re-interpretation of Loar, Mary’s inability to infer WIL like from neural knowledge is not a consequence of the fact the phenomenal concepts are neutral, but rather, it is a consequence of the fact that such concepts can be possessed only by someone who has the qualitative states in question. Those qualitative states are essential components of the concepts. With that concession, Max Black returns to the scene. I argued earlier that, if knowing WIL to see red is knowing a property of red experiences, it follows that red experiences have a property that is not physical. We are back to square one.

## 2.7 Acquaintance

The predicament can be summarized as follows. We saw three versions of the view that Mary learns a new mode of presentation on her release. According to the first, modes are individuated by properties. The problem here is that those properties would have to be non-physical, because Mary knows all physical properties before her release and she can't use those to explain the alleged non-physical ones. According to the second mode of presentation reply, modes are individuated by causal roles. The problem here is that causal roles seem too abstract and hence too easy for Mary to acquire before her release. The solution is to pin down those roles by defining them in terms of causal relations to experiences. This is an aspect of the third mode of presentation view, which emphasizes recognitional concepts. The problem with this move is that, if our phenomenal experiences are implicated in our phenomenal concepts, then those concepts present their referents via phenomenal properties, and that concession makes the third mode of presentation view into a version of the first. It looks like the physicalist might have to admit defeat.

In addition to these specific problems, there are two general worry about the mode of presentation approach. There does seem to be any mode—any mental representation—that Mary couldn't have before learning WIL. Earlier, I introduced the case of Subliminal Mary. She has a representation in her visual system induced by a very brief presentation of a red stimulus followed by a mask. Subliminal Mary has a visual mode of presentation representing red, but she does not know WIL. In response, friends of the mode of presentation strategy might point out that Mary does not have a phenomenal concept. She cannot think about her red percepts. Let's imagine, however, that Subliminal Mary is trained to make accurate forced choice judgments about what colors she has seen. She is cued to guess every time she sees the mask appear. We can call her Supersubliminal Mary after training. Now Mary can think about her red percepts. She can even recognize them. Presumably she has conceptual modes of presentation representing red. Yet, she doesn't yet know what red is like. This case *may* turn out to be impossible, but it isn't obviously impossible. If it is possible, it follows that having a mode of presentation of a special kind may not be sufficient for knowing WIL.

Fortunately, there is a promising response strategy that I have not yet considered. Rather than saying that knowing WIL is a matter of possessing a mode of presentation, one can say it is a matter of having *a special kind of epistemic access* to that mode. On this approach, Mary doesn't learn any new properties, but rather she acquires a new way of knowing. Lewis and Nemirow's ability theory pushed in this direction. They distinguished knowing how from knowing that. But this strategy failed, because the abilities they consider are neither necessary nor sufficient for knowing WIL. But there is another distinction to be drawn between two kinds of knowing. We can distinguish knowledge by acquaintance and knowledge by description (Russell, \*\*\*: chapter, \*). Many languages mark this difference lexically. There is one epistemic verb that expresses knowledge by acquaintance or familiarity (e.g., *conoscere, conocer, connaitre, kennen*), and another that expresses master of facts (e.g., *sapere, ser, savoir, wissen*). Some authors have suggested that, after leaving her room, Mary merely acquires

acquaintance with properties she already knew by description (Horgan, \*\*\*; Conee, \*\*\*; see also Churchland, \*\*\*, on two kinds of knowing).

This proposal can be used to refute the third premise of the knowledge argument. Jackson assumes that, if Mary cannot deduce knowledge of WIL from physical knowledge, then WIL is not physical. But this rests on an equivocation. Knowledge of WIL is not knowledge by description; it is acquaintance. One cannot infer acquaintance from descriptive knowledge. Acquaintance occurs only when one has encountered something, and describing something is not the same as encountering it. The fact that we cannot infer acquaintance from description has no important ontological implications. You can describe Timbuktu to me, but I can't become acquainted with Timbuktu without going there; it does not follow that Timbuktu is not physical. Likewise, that fact that Mary's textbooks do now acquaint her with color experiences does not entail that those experiences are not physical.

This move is seductive but problematic. The difficulty is that the gulf between acquaintance and description usually isn't very large. Usually, one can learn so much information by description that acquaintance yields no surprises. Suppose I tell you about my friend Henrietta. I tell you all her personality, her biography, and her appearance in excruciating detail. Perhaps, we have a lot of time on our hands, so I tell you everything there is to know about these facts. Of course, you are not thereby acquainted with Henrietta. You don't know her. Description does not entail acquaintance. Now suppose I introduce the two of you. You now acquire a new form of knowledge: knowledge by acquaintance. But, I submit, you won't be surprised when you meet her. If my descriptions have been thorough, Henrietta will be just as you would have expected.

The point can be made by saying that acquaintance is an externalist epistemic construct. When you shift from description to acquaintance, the only thing that must be added is some kind of *encounter* with the object of knowledge. The internalist (i.e., mentally represented) information can be entirely descriptive. Encountering something can increase your stock of internal information, but that information is not necessarily qualitatively different than information you could have gotten by description without acquaintance. In other words, in many cases, a person who is *not* acquainted with something may have internal states that are indistinguishable with someone who *is* acquainted with that thing. Acquaintance is too epistemically modest to account for Mary's surprise when she leaves the room.

In response, the proponent of the acquaintance strategy might argue that knowing WIL involves a special kind of acquaintance, which is much more than description-plus-encountering. But this move renders the acquaintance proposal empty. What is the special kind of acquaintance that Mary acquires? Well, it's phenomenal acquaintance of course. Mary's surprise on leaving the room is not merely a consequence of her becoming acquainted with something; the appeal to acquaintance does no explanatory work. Her surprise is rather a consequence of having phenomenal experiences, and the physicalist owes us an explanation of why we can't attain an epistemic state that is internally indistinguishable from these by description alone.

Despite these complaints, something seems right about the acquaintance approach. By focusing on how representations are accessed rather than focusing on the representations themselves, one can overcome some of the objections facing mode of

presentation replies to the knowledge argument. What we need is a substantive account of the special epistemic access relation, rather than the unhelpful label “acquaintance.” We need to explain what our *direct epistemic access* to experiences consists in.

Coming up with a substantive account of direct access is not going to be easy. There is a challenging puzzle concerning direct access. In cases of indirect epistemic access, we can explain two things: what we know about the entity accessed *and what we don't know*. I have good indirect epistemic access to my shoes, but it is not good enough to tell me about the molecules comprising them. I can't see molecules as such. This failing is explained by appeal to a limitation in my representational acuity. But consider direct epistemic access. If we are in direct epistemic contact with the thing itself, then there should be no barrier to accessing any of its essential properties. This is potentially embarrassing for the physicalist. Physicalists claim that experiences are token-identical to brain states. If we have direct access to brain states, and direct access makes every essential property available, then we should have direct access to the brainy properties of our experience (such as the morphology and firing pattern of their constituent cells). On the face of it, we have no such access.

This can be brought out by considering the Reverse Knowledge Argument. Mental Mary is Mary's sister. She is just like us. She knows nothing about the brain, but she knows a lot about experience. She is great at introspecting, and, when she experiences red she likes to focus on phenomenal qualities of that experience. She knows everything there is to know about the phenomenology of seeing red. Yet, she does not know that red experiences are brain states. If she were to read about the underlying cellular activity, she would be extremely surprised. If knowing WIL were a matter of having direct epistemic access to brain states, then one might expect Mental Mary to be able to discern the brainy properties of her experiences. She should be able to read them off, in much the way that we can describe the brush strokes in a painting by Van Gogh. Her inability to do that places some pressure on the direct access thesis and cries out for explanation.

This puzzle about direct access thesis provides another desideratum on a solution the knowledge argument:

**The Mental Mary Condition:** An adequate account should include a substantive theory of direct access that explains why we can't read brainy properties off of our phenomenal states.

I don't think these desiderata are impossible to meet, but I claim that none of the standard replies to the knowledge argument meets them all. Each desideratum has been generated by considering an objection to one of the standard replies. In the remaining section, I will propose a response to the knowledge argument that satisfies all of the desiderata.

### 3. Maintaining Mary

#### 3.1 Maintained Availability

Thus far, I have been considering a variety of different theoretical posits that might be used to explain why Mary can't infer WIL from her knowledge of physical facts. We have talked about imagination, mnemonic abilities, modes of presentation, recognitional concepts, and acquaintance. Almost all of these proposals have been generated in a philosopher's armchair with no effort to find empirical support. That is peculiar. After all, the posits in question are substantive. The defenders of these theories are not merely making verbal moves; they are speculating about the architecture of the mind. Armchair speculation is a good way to generate testable theories, but it should be viewed as a starting place, not the final word. We have seen that some of these starting places are non-starters. None of these posits would be able to adequately answer the knowledge argument, even if they could be empirically confirmed. For example, the ability view cannot explain learning what it's like in the absence of the proposed abilities; the mode of presentation view tends to proliferate properties; and the acquaintance view cannot by itself explain why Mary is incapable of knowing WIL in the absence of actual contact with phenomenal states. Thus, prevailing proposals are doubly flawed: they are empirically unfettered, and they are not able to satisfy all the desiderata on an adequate solution to the knowledge argument. Perhaps, if we begin by looking for an empirically motivated theory, we will end up with a better explanation of why Mary doesn't know WIL.

In earlier chapters, I presented an empirically driven theory of consciousness. According to that theory conscious states are AIRs, attended intermediate-level representations. Consciousness arises when an intermediate-level representation is made available for processing in working memory through attentional modulation. Let's assume that this theory is correct, and ask how Mary learns what it's like to see red after leaving her room.

Presumably, Mary's first encounter with a red object would, if her parvocellular system had not atrophied, cause activation in V4 neurons that are responsive to wavelengths between 600 and 700 nanometers. If the stimulus were presented long enough for her attention systems to engage, the V4 response would reach thresholds high enough to propagate forward, sending afferent to brain structures that can temporarily store representations that carry spectral information. As soon as the V4 cells broadcast to systems higher in pathway—as soon as their responses become *available*—Mary has a conscious experience of red. That experience is like something.

Strictly speaking, however, Mary might not yet *know* what its like. Knowing an experience seems to require having the experience, but having the experience does not suffice. Intuitively, there could be creatures for which it is like something but who do not know what it is like. Some readers may want to insist that having an experience constitutes a form of knowing WIL in and of itself. Perhaps. I don't want to rule out a primitive meaning of "knowledge" here. It must be conceded, however, that Mary ultimately learns something more. She does not merely have an experience. She learns that red is like *this*. That form of demonstrative knowledge can play a role in her deliberations about phenomenal states.

What more is needed? Fans of recognitional concepts think Mary needs an internal state that represents her red experience and can be used to recognize it. I find this implausible. We can ostend experiences that we could never again recognize. Examples are easy to generate: “look at that unusual shade of blue,” “I have this weird queasy feeling that I’ve never had before,” “I wouldn’t recognize a piccolo if I heard one, but during the concept I thought, now *that’s* a lovely sound.” Thus, demonstrative reference to experiential states must be less demanding than conceptual reference. But how does it work? Here, again, we should look to the science.

In chapter 3, I distinguished availability to working memory from encoding in working memory, though the latter is the natural sequelae of the former. To have thoughts about experiences, encoding must take place. But what is working memory encoding? A prevailing view in cognitive neuroscience is that working memory is not really a storehouse, but a collection of “executive” processes. For example, memory can be used to rotate mental images, draw inferences, make action decisions, or orchestrate strategic deployments of attention. But, more fundamentally, working memory can *maintain* activity in sensory cortices (Pertides, 1996; D’Esposito et al., 1998; Glahn et al., 2001). It is by maintaining representations that working memory can keep a stimulus in mind after it has been taken away. Suppose you read a number in a phone book, and then hold it in your head on route to the phone. Working memory is keeping it there. Or suppose you see a crime, and keep the assailant’s face in mind until the police come. Working memory does that. Working memory can also maintain focus on a stimulus that remains present, as when we track the ball in tennis. Here, we might imagine a feedback loop from vision to attention to working memory, and then from working memory via attention back to vision. For present purposes, I want to concentrate on the idea of maintenance. As a first stab, I propose that a working memory encoding is an internal state that allows us to maintain a state in our perceptual systems. When Mary thinks “red is like *this*,” she is holding a red representation in her visual system. As with the tennis ball, we might imagine a feedback loop. The red representation in V4 becomes available via attention to working memory, and then a working memory state is generated, which allows Mary to retain that V4 representation for a brief period of time. Mary has access to what red is like as soon as the V4 state becomes available, and she can think about what it’s like as soon as she is in a position to maintain that V4 state.

This story can explain why Supersubliminal Mary does not know what it’s like to see red. Having a red representation in one’s visual system is not sufficient for having a red experience. One might even learn to recognize unconscious visual representations and make discrimination judgments. Having qualitative experience requires more than a representation; it requires that the representation be available to working memory. Knowing what representations are like, then involved maintaining them in working memory. Supersubliminal Mary can do neither. I would predict that if Supersubliminal Mary gained working memory access to her color representations, they would no longer be subliminal; they would rise above the threshold of consciousness.

Now let’s consider this proposal in the context of the knowledge argument. Prior to leaving the room, Mary knows about the AIR theory. She knows all the facts about what goes on in the brain when people see red. Nevertheless, these states have never occurred in her brain. V4 representations of red have never been made available to her. Since such events are token identical with the experience of red, it follows that she has

never had an experience of red. Now after leaving the room, she has that experience. Mary is also able, after leaving the room, to have the thought expressed by “red is like this.” That thought becomes possible when her color experience is maintained by the executive processes in her working memory system. Mary learns something new when she has this thought. It would be misleading to say she learns a new fact, because she already new about all the states of affairs that have to occur for red experiences to arise. It would also be misleading to simply say that Mary has a new mental representation. If Mary had been exposed to red subliminally, she might have already had representations in V4. Instead, Mary gains access to a representation; the representation becomes available for thought and deliberation, and she maintains it through executive control in working memory. Learning WIL might best be described as attaining a new relation to perceptual states. Knowing WIL is maintaining availability.

The third premise of the knowledge argument says, if physicalism is true, then Mary should be able to derive WIL from her knowledge of physical and functional properties of the brain. If the AIR theory is right, then this premise is false. Knowing WIL is maintaining availability, and knowledge of the brain is not sufficient to put one in that position. The third premise of the argument says physicalism is committed to the derivability of phenomenal knowledge, but, if such knowledge is constituted by standing in certain relations to one’s brain states, the derivability assumption is false. Physicalists should reject P3, and the AIR theory explains why.

### *3.2 The Desiderata*

The maintained availability solution to the knowledge argument fares well in comparison to other proposals. So see that, let’s consider how it satisfies each of the desiderata that were adduced earlier.

**The Imagination Condition:** An adequate account of knowing WIL should explain why physical descriptions do not always allow us to imagine WIL.

Why can’t Mary imagine what red is like after reading everything about the brain? Because such knowledge would not allow her to generate perceptual states of the right kind, much less maintain them. To know what red is like, she might have a visual state of the kind that registers the presence of red, but she cannot generate such a state by act of will. If Mary gained uncanny control over her visual system, she might be able to imagine things she had never seen, but her imaginative abilities will always be limited by her neural hardware. She cannot imagine what it’s like to be a bat.

**The Solipsism Condition:** On any adequate account, knowing WIL cannot (merely) be a matter of knowing what external features of mental states represent.

As a matter of contingent fact we can individuate phenomenal states by their referents. Throughout, I have been talking about an experience of red with the assumption that red is a spectral property of the external world. But we cannot rely on this method of individuation to heavily, because, for example, different sense modality

can represent the same content in phenomenally different ways. How do we know that we are seeing a spatial location as opposed to hearing one? How do we know that we are *tasting* sweetness as opposed to *smelling* it? The answer suggested by the AIR theory is that every qualitative state is a perceptual representation. We can distinguish these representations by their contents, but also by their internal roles. Think about this from a solipsistic point of view. Suppose I don't know what my red experience represents, because I don't know what the property of being red consists in. Yet, I know how red differs from blue. I speculate that this knowledge has to do with maintenance. Maintenance involves the selective deployment of processes that entrain sensory systems. I can look at a field of colors and maintain focus on the blue while ignoring the red. My ability to selectively enhance and maintain different percepts constitutes a kind of discriminative capacity. I have that capacity regardless of what my percepts refer to. I'd have it if I were a brain in a vat.

**The Disability Condition:** On an adequate account, knowing WIL cannot require the capacity for imagination, recognition, or recall.

On the maintained availability view, Mary need not acquire the abilities emphasized by Lewis and Nemirow in their response to the knowledge argument. Suppose she sees a strawberry for the first time. She can maintain the percept while staring at the strawberry, but she might not be able to recall or imagine its hue on a future occasion, and her capacity to recognize that hue on her next encounter might be very low.

**The No Property Condition:** An adequate account of knowing WIL cannot assume that Mary represents her experience of red via a property.

How does Mary represent red on the present view? One possibility is that she represents red by means of her working memory encoding. I don't think that's very likely. Working memory encodings are executive processes that maintain activity in sensory systems. We don't need to think of them as representations at all, and if we do think of them as representations, then they are best regarded as instructions on how to maintain a red state. Working memory encodings encode procedural knowledge—a kind of know-how. It would certainly be misleading to say that Mary represents her red experience simply by *representing* how to maintain it; what matters is that she *can* maintain it. In this respect, the maintained availability view is like the ability view. On the other hand, we can also ascribe a demonstrative representation to Mary: the representation expressed when she says, “red is like *this*.” The demonstrative corresponds not just to her procedural knowledge but to that knowledge as exercised and the resultant maintained availability of Mary's red percept. We can think of maintained-red-percept as a representation—but of a very unusual sort. The maintained-red-percept does not *refer* to red; rather it *exemplifies* red. It is like the color sample in Wittgenstein's dictionary. I am inclined to say that Mary does not represent her red experience at all when she forms the demonstrative thought. She just has a red experience. Consequently there is no worry about her representing red by means of a non-physical property.

**The Independence Condition:** On an adequate theory, phenomenal concepts cannot be individuated by relations to other phenomenal concepts (i.e., holistically).

There is a sense in which the maintained availability theory is a causal role theory of phenomenal concepts. The demonstrative concept expressed by “this red” involves a maintenance relation between working memory and perceptual states, and maintenance is a causal notion. But I have not suggested that colors get their individual character by causal relations with each other. So I am not committed to holism about qualitative character of color experiences. I think each color (and each experienced perceptual representation) has a role that can be described as vertical rather than horizontal. A horizontal role relates each experience to each other experience of a certain type. Color experiences have horizontal roles, but those don’t account for distinct color qualities. Vertical roles are roles between a color experience and the states that precede and follow them in an information pathway. These roles are more important. I suggested earlier that our capacity to distinguish experiences that have the same content but different feels is connected to the fact that these experiences are in different pathways and they are maintained by processes that can sustain activity from earlier states of processing. These ideas contribute to an understanding of why a person with unilateral color blindness can experience some of the same colors in both eyes.

**The Phenomenal Constituent Condition:** An account of knowing WIL to have an experience should assume that the knowledge is constituted in part by the experience itself.

The maintained availability account satisfies this requirement. Knowing WIL is not merely a matter of having a working memory encoding. It is a matter of maintaining a perceptual state, and that perceptual state is the correlate of the conscious experience. Consequently, one cannot know WIL without having the experience. In this respect, phenomenal concepts are not neutral (like Loar’s recognitional concepts on one interpretation), but neither are they descriptive. They contain experiences but they don’t represent experiences.

**The Mental Mary Condition:** An adequate account should include a substantive theory of direct access that explains why we can’t read brainy properties off of our phenomenal states.

I will spend a bit more time on this problem, because it has received surprisingly little attention in the literature. Everyone is worried about the fact that can’t infer the phenomenal qualities from neural descriptions, but comparatively few are worried about the fact that we can’t infer neural descriptions from phenomenal qualities. Perhaps the neglect stems from the fact that contributors to the consciousness literature assume that that this problem has an easy solution. But a quick review of knee-jerk responses suggests that the problem may be harder than it looks:

*Proposal 1:* Phenomenal states are multiply realizable, and one cannot infer a

realizer from a multiply realized state, because there is an open-ended range of realizers.

*Reply:* It's not obvious that qualia are multiply realized, and in any case, most physicalists accept that token-identity theory. An adequate reply should explain why direct epistemic access to a token phenomenal state does not reveal its braininess.

*Proposal 2:* We have direct access to brainy properties as such, but we cannot "read them off" because they are ineffable, non-conceptual, or otherwise inaccessible to cognitive mechanisms that we use to report our phenomenal states.

*Reply:* This strikes me as a peculiar claim. After all, phenomenal states seem to be vividly accessible. We can describe them in more detail than any unconscious mental states. When we focus on a conscious experience, we have no difficulty reflecting on it, reasoning about it, or reporting it.

*Proposal 3:* We have direct access to brainy properties as such, but we don't describe experiences that way because doing so requires possession of a neuroscientific vocabulary.

*Reply:* If this were the case, then we should at least be able to describe our experiences using geometrical vocabulary that captures corresponds to neural structures. That is, we should be able to discern that our experiences decompose into roundish parts that have long branching lines stemming off of them. Later, when we learn the words "cell body" and "dendrites," we should say: "Oh, that's what those things are!" Of course, we don't do this.

*Proposal 4:* Phenomenal states are diaphanous. They represent external features of the world, and when we access them, we are aware of those features.

*Reply:* I argued against representationalism above. I think both representational content and apparent diaphanousness are contingent features of phenomenal states. There are cases on record of individuals who had sight restored after a lifetime of blindness (von Senden, 1932; Gregory & Wallace, 2001). Those individuals have to initially interpret the meaning of visual stimuli. Gregory & Wallace describe one case in which a man sees a face for the first time infers that it is a face from the fact that he hears it speaking. It is plausible that when these individuals (or human infants) first see, they experience visual percepts as purely internal events, not as representing anything.

*Proposal 5:* One might think the problem of Mental Mary is based on an

unwarranted allegiance to the act-object theory of perception. According to that theory, perception involves an act of awareness directed towards an object that serves as an intermediary between mind and world. On this view, having a phenomenal experience is like staring at a mental painting. If the act-object theory were right, we should be able to simply describe the medium in which our experiences are painted. But the act-object theory is false. Phenomenal states are not objects viewed by a mental eye, but vehicles through which we view the external world.

*Reply:* My answer parallels what I said in response to representationalism. When we have phenomenal experiences we are, in a sense, aware of *them*, not (merely) of what they represent. Perhaps it is a mistake to say we view our phenomenal experiences with an inner eye, but neither are they transparent windows onto the world. We experience and reflect on their properties. In any case, it's not obvious that collapsing act and object would help here. On the act-object view, experiences are inner paintings viewed by an inner eye. On some more sophisticated views—such as Frith's (1949-50) "Percept Theory"—experiences are like inner paintings with no viewers. But, even without viewers, paintings have intrinsic properties, and these are certainly available to information-processing systems further down stream.

I don't presume that these replies are decisive. Some version of one of these proposals might be salvageable. I offer this brief survey to indicate that the solution to the Reverse Knowledge Argument is not totally obvious. In any case, I want to propose a somewhat different response.

I want to begin by picking up on the act-object theory of perception, which has an important kernel of truth. On the AIR theory, consciousness arises when a perceptual representation (an object) is made available to working memory via attention (an act of awareness). Structurally, the view conforms to old-fashioned act-object theories, but there are also important differences. Old-fashioned act-object theorists believed in sense-data (e.g., Russell, 1912; Broad, 1923; Price, 1932). Sense-data were presumed to be non-physical entities that have no intentional content, and literally have properties like redness, roundness, and so on. Perceptual representations, on my view, are physically realized states that ordinarily represent features of the world. In these respects, they are quite unlike sense-data. I also part from the old-fashioned act-object theory in another way. Sense-data enthusiasts claimed that the act of perception involves the assignment of meaning to otherwise meaningless sense-data. Perceptual acts are interpretive acts. On that view, we inspect our sense-data, and then figure out what they tell us about the outside world. On the view I favor, the act of awareness is not an interpretive act; it is the act of maintaining availability. The working memory encodings that do this work are not, I have suggested, representations of perceptual states; they are processes that keep perceptual representations active and accessible. This is an inversion of sense-data theory. On that approach, the objects of perception are contentless, and the acts confer content; on my approach, the objects of perception are (ordinarily) contentful, and the acts are contentless. With this inversion, we can begin to discern a solution to the Mental Mary problem.

The Mental Mary problem arises because we seem to be able to inspect our perceptual states. If we are inspecting them, and if inspection is direct, then we should be able to identify their brainy features. But I think inspection is an illusion. What seems like the inspection of a perceptual state is not a semantic or epistemic act, but rather an act of selective entrainment. We maintain one portion of a percept, and then another, and then another, making each available for further processing in turn. What seems like inspecting is just a form of moving about. Knowing the features of our perceptual states is like knowing our way about.

When we perceive a red surface, we form a perceptual representation of that surface, and this results in an experience of the phenomenal quality we call “redness.” What is it to know what the quality of redness is like? It is to maintain that representation. What is the phenomenal quality of redness? It is the one I attend to in this way. Knowing what redness is like is a kind of procedural knowledge. It’s a matter of knowing how to maintain or focus in on features of a perceptual representation. If I show you a field of colored shapes and say, “Look at the red ones,” you will respond by using a stored perceptual template to search for matching features in your perceptual representation, and those matches will lead you to attend selectively to the red shapes. Top-down visual search engages the mechanisms of maintenance. You seek out features by exercising your procedural knowledge—by finding your way about. Once those features are found you can maintain their availability by fixing attention.

Let’s suppose this picture is right. We can then explain why Mental Mary cannot discover brainy features in her perceptual states. Mental Mary is directly aware of her perceptual state, which is to say her perceptual state is made conscious through an act of availability. Her perceptual state is (at least token identical to) a brain state. So Mental Mary is directly aware of a brain state. But, her grasp of the “features” in that state is not a matter of representing them; it is a matter of maintaining them. Being aware of a perceptual representation is not a matter of forming another representation—or interpretation. So, Mary cannot read off the brainy properties *or any other properties* of her experience. At least not yet.

At a subsequent stage of processing, Mary can reflect on what the components of her experience signify; she can engage in acts of interpretation. But these acts are post-experiential. They are not part of the perceptual process or components of the perceptual experience. In information processing terms, we can imagine that the maintained perceptual representation sends signals to perceptual memory systems and systems that allow verbal reports. These systems do represent the experience, but not by direct access. They contain categorical representations that register similarities across experiences, and they can be used to classify and label what we are seeing. But these representations are not conscious, and they do not represent neural features. Some of them may be like Smart’s topic neutral representation; they inform us that we are seeing something of the kind that we experience when we look at a ripe tomato in good light. We may become aware of these interpretations through subvocal speech or mental imagery, but they do most of their work behind the scenes.

So we can think of perception as proceeding in stages. A stimulus is presented causing a perceptual representation. That representation becomes available to working memory through top-down or bottom-up attention. At this point, there is something it is like to have this representation. Once in working memory we can maintain the

representation, by using executive faculties to sustain the flow from perceptual systems into working memory. This involves a selective deployment of attention. Access to WIL occurs at this stage. Access to what its like consists in holding the representation there and, in some cases, moving around the focus of attention. Access to WIL is not a matter of reading-off features. Reading off can be achieved at a post-experiential stage in which we label or otherwise classify component of the experience. The labeling stage does not have direct access to experience. It works by *representing* the experience. Mechanisms of labeling detect similarities, but the labels do not represent brainy features, nor, *a fortiori*, do they make such features available to consciousness.

This solution to the Mental Mary problem may seem similar to the fifth proposal that I canvassed above. In line with that proposal, I have now suggested that conscious states are not inspected objects in a mental gallery. But, unlike proposal five, I am not suggesting that we drop and act-object theory of perception. Perception is a matter of performing acts of awareness on mental objects. The major fault of the traditional act-object theory is that the acts of awareness were regarded as interpretive acts. If we drop that assumption, there is no longer any pressure to think that we should be able to inspect the medium of our percepts. We move about through our mental paintings; we don't inspect them.

In the previous subsection, I suggested that the maintained availability story can be used to show that the knowledge argument is unsound. In this section, I argued that it also satisfies the desiderata on an adequate solution to the knowledge argument. Other replies to Jackson fail on one or another desideratum. I maintain that maintained availability is a better response.

### *3.3 What Sort of Solution Is This?*

In closing, it will be helpful to say a little more about how the maintained availability story relates to other theories. I described a number of specific proposals for rejecting the third premise of the knowledge argument, but these partition into basically three strategies. First, there is the ability strategy. On this approach, Mary does not learn any new facts, but she learns new abilities. Then there is the mode of presentation strategy. Here, the idea is that Mary does not learn new facts (construed as states of affairs), but instead learns old facts using a kind of concept or representation that was unavailable to her before. And finally, there are acquaintance theories, according to which Mary does not necessarily even acquire a new representation; she just gains a new form of epistemic access. Where does the maintained availability picture fit in?

I think each strategy gets something right. On the maintained availability theory, Mary does gain a new ability: the ability to selectively maintain her perceptual states. I said much about the role of procedural knowledge (know-how) in responding to the Mental Mary problem.

Mary can also be said to acquire a new demonstrative concept when she leaves the room. I have insisted that working memory encodings that sustain perceptual activity are not representations, and, thus, I would not call them concepts. But the composite of a working memory encodings plus the perceptual activity that they maintain can qualify as conceptual. On one definition, a concept is a mental representation that can be a

component of a thought. The maintained perceptual state can be a component of a thought. It is what Mary expresses by the demonstrative when she says, “Red is like *this*.” The maintained perceptual state can be regarded as a representation (where representations are defined symbol tokens), but I cautioned that it does not necessarily represent. The maintained perceptual state exemplifies phenomenal qualities; it does not denote them.

In addition, Mary can be said to acquire a new epistemic relation to her perceptual states when she leaves the room. If Mary had been exposed to subliminal presentations of colors, she might have formed red representation. On leaving, she gains direct access to these, because they become available to working memory. The availability account offer a substantive analysis of what some people might mean by “acquaintance.”

In sum, the maintained availability story has something in common with each of the prevailing strategies for responding to the knowledge argument. But, because it borrows from all, it overcomes the problems of each. The standard ability account cannot explain the intuition that Mary forms a new thought (“Red is like this”); it could avoid this problem by appealing to representations. The standard mode of presentation account tends to proliferate properties by proliferating representations; it could avoid this problem by introducing a special epistemic access relation to existing representations. The acquaintance account does not provide a substantive analysis of the required epistemic access relations; it could avoid this trouble by availing itself of access abilities. Each approach needs the others. The maintained availability account is a way to bring these resources together into a unified whole.