Inexpressible Truths and the Allure of the Knowledge Argument*

Benj Hellie†

May 27, 2003

Dualism is a perpetually seductive doctrine; the Knowledge Argument for dualism (Jackson, 1982) a particularly alluring source of support for the doctrine. Jackson advocated the soundness of the argument for nearly two decades before changing his mind; I and many of my comrades, before we became sophisticated, found the allure of the argument as Jackson presents it hard to avoid; many other philosophers doubtless have had this experience as well. And, as Stoljar and Nagasawa (2003, §§1–2) detail, related ideas have cropped up in the literature throughout the century.

In this paper, I argue that the root of this allure lies in the Knowledge Argument's involvement with inexpressible concepts. I begin with a discussion of the meaning of a certain version of the Knowledge Argument: in §1, I describe this version and argue that it is formally valid; in §2, I analyze the vexing sentence 'Mary doesn't know what it's like to see a red thing', and argue for the existence of a certain sort of inexpressible concept. In §3, I assess the Knowledge Argument for soundness. I argue that no fallacy is involved; rather, the argument is apparently sound. Whether its conclusion should ultimately be accepted hangs on the status of a certain attractive doctrine concerning the ontological impact of inexpressible

*Thanks for discussion of these issues to many over the years. I'd especially like to thank Dave Chalmers, Mike Fara, Gail Fine, Sally McConnell-Ginet, Peter Ludlow, Kieran Setiya, and Jessica Wilson. Harold Hodes, Sydney Shoemaker, Daniel Stoljar, and Zoltán Szabó provided helpful written comments; Sydney's and Zoltán's comments moved me to fundamentally rethink my approach to sizeable portions of the paper, and Daniel's comments on several drafts were impressive both in quantity and in quality. Thanks also to three classes of students in Intro Phil Mind for serving as test subjects.

†Sage School of Philosophy, Cornell University; benj.hellie@cornell.edu

concepts.

* * *

Autobiographical disclosure: I am not a dualist, and I do not intend this paper to establish dualism. Rather, I hope that dualism is not correct: I love desert landscapes, naturalism, and structural explanation as much as the next philosopher. However, I suspect that physicalist philosophers have been overhasty in their treatment of the Knowledge Argument, being quick to mistakenly accuse it of this or that fallacy. As a result, the real threat posed by the Knowledge Argument has been missed. My opinion is that this threat won't go away until it is exposed and met head-on; and that it won't be possible to expose this threat unless the argument is treated with the sympathy its allure has earned it.

1 The Knowledge Argument is Valid

Jackson (1986, p. 293) explicitly presents the following argument concerning the "Black-and-White Mary Scenario":[1]

1_J Mary (before her release) knows everything physical there is to know about other people.

2_J Mary (before her release) does not know everything there is to know about other people (because she learns something about them on being released).

C_J Therefore, there are truths about other people (and herself) which escape the physicalist story.

This argument can be weakened and tidied up somewhat to yield the following argument, which will be the focus of this paper:

At the time immediately before Mary's release,

1 Mary knows all the physical truths.

2 Mary does not know what it is like to see a red thing.

C There is a truth which is not physical.

The argument (1), (2); (C) is superior to (1_J), (2_J); (C_J) for two reasons: first, (2) is closer to our immediate intuition about the Black-and-White Mary Scenario than is (2_J): while our acceptance of (2) is presumably the source of the plausibility of (2_J), some have accepted (2) while denying that the failure to know what something is like is not the failure to know a truth (Lewis, 1988; Bigelow and Pargetter, 1990) (and although (1) is rather stronger than (1_J), the additional strength does not seem in any way to effect the dialectical situation). And second, (C) is slightly clearer than (C_J): saying that a truth "escapes the physicalist story" is clearly intended to be a fancy way of calling it non-physical (compare Jackson 1986, §I: "[b]ut she knew all the physical facts about them all along; hence, what she did not know until her release is not a physical fact about their experiences").

I can't find an argument with weaker premisses than those of (1), (2); (C) that remains true to Jackson's intentions. For this reason, though there are many things which may deserve the title 'the Knowledge Argument', I will (somewhat stipulatively) reserve the term to denote (1), (2); (C).

If we wish to explain the allure of the Knowledge Argument, we must establish what we pretheoretically take the Knowledge Argument to mean. It will be helpful to begin by establishing the most abstract, formal, logical properties of the Knowledge Argument. The logical form of the Knowledge Argument can be represented as follows:[2]

$\mathbf{1_{lf}}$ $(\forall t)(\Phi t \supset Kmt)$
$\mathbf{2_{lf}}$ $(\exists t)(At \wedge \neg Kmt)$
$\mathbf{C_{lf}}$ $(\exists t)(\neg \Phi t)$

The non-logical expressions in this (valid) form are used in the following abbreviations, where small Greek letters are used as syntactic variables and corners represent quasi-quotation: 'Mary' $\Longrightarrow$ 'm'; $\ulcorner \tau$ is a physical truth$\urcorner \Longrightarrow \ulcorner \Phi\tau\urcorner$; $\ulcorner \mu$ knows $\tau\urcorner \Longrightarrow \ulcorner K\mu\tau\urcorner$; and, to put it approximately pending further clarification, $\ulcorner \tau$ "answers" the question 'what is it like to see a red thing?'$\urcorner \Longrightarrow \ulcorner A\tau\urcorner$. The next two sections will be primarily concerned with explicating (2_{lf}) and justifying its assignment as logical form to (2).

It's worth pausing a moment to consider the frequently encountered complaint that the Knowledge Argument is invalid due to the phenomenon of the "referential opacity" of 'know': where ρ and σ are singular terms, from $\ulcorner\mu$ knows that $\cdots\rho\cdots\urcorner$ and $\ulcorner\mu$ does not know that $\cdots\sigma\cdots\urcorner$, $\ulcorner\rho \neq \sigma\urcorner$ does not follow. If this complaint is directed at the Knowledge Argument as I am understanding it, it is clearly at odds with my claim that the Knowledge Argument has the logical form (1_{lf}), (2_{lf}); (C_{lf}).

Fortunately, this complaint can be easily parried. Jackson is clearly aware of the opacity concern, and intends to present an argument that is not invalidated by it. In the 1986 paper he insists that "the intensionality of knowledge is not to the point. The argument does not rest on falsely assuming that, if S knows that a is F and $a = b$, then S knows that b is F. It is concerned with the nature of Mary's total body of knowledge before she is released: is it complete, or do some truths escape it?" (§I); and he replies to an early expression of the opacity concern as follows: "What is immediately to the point is not the kind, manner, or type of knowledge Mary has, but *what* she knows. What she knows beforehand is ex hypothesi everything physical there is to know, but is it everything there is to know? That is the crucial question" (§II).

Charity thus recommends against assigning an invalid logical form to the argument, if possible. And this *is* possible: the standard ordinary language uses of 'truth' and 'know' license the principle that if someone knows that p but does not know that q, then the truth that p is numerically distinct from the truth that q: from $\ulcorner\mu$ knows the truth that $\pi\urcorner$ and $\ulcorner\mu$ does not know the truth that $\rho\urcorner$, $\ulcorner$the truth that $\pi \neq$ the truth that $\rho\urcorner$ *does* follow. In this respect, truths are like ("Fregean") propositions. For instance, if someone knows that Hesperus is Hesperus, but doesn't know that Hesperus is Phosphorus, then the truth that Hesperus is Hesperus is numerically distinct from the truth that Hesperus is Phosphorus. 'Know' is not opaque, that is, with respect to terms denoting truths; truths are, in a slogan, the objects of knowledge.[3] There may be some distinction between truths and true propositions; but for my purposes here, this distinction won't matter, so I will ignore it, treating truths and true propositions alike.

2 'What It's Like' and Inexpressibility

I say that the logical form of (2) 'Mary does not know what it's like to see a red thing' can be represented (approximately) by (2_{lf}) '$(\exists t)(t$ "answers" the question 'what is it like to see a red thing?' $\wedge \neg$(Mary knows t))'. Why say this? And what does it mean?

I begin with a methodological discussion of the appropriate way to investigate the logical form of (2). So far as I can tell, the dominant view among philosophers is that uses of the frame 'what $\cdots$ is like' in the context of the philosophy of consciousness are not "normal": i.e., they are not pieces of fully literal ordinary language, but rather involve some use of (for instance) metaphor or idiom or jargon or code. While this opinion is rarely stated explicitly, it is frequently asserted in conversation and is often insinuated in writing—e.g., by putting a completion of the frame in italics or in scare quotes, or by running a completion of the frame together as a single word, or by using such a completion ungrammatically, or by attributing the frame to Nagel, or by calling it a "stock phrase", or by using it in some other self-conscious way to indicate that the frame is somehow not wholly in order. I would guess that most philosophical books or articles on consciousness published in the last decade do this.

Lewis (1995, p. 326) stands out for providing an explicit argument that, in this context, the frame has a "special technical sense". He argues that "[y]ou can say what it's like to taste New Zealand beer by saying what experience you have when you do, namely a sweet taste. But you can't say what it's like to have a sweet taste in the parallel way, namely by saying that when you do, you have a sweet taste!" While this passage is somewhat obscure, I take Lewis to be arguing that, in literal, ordinary use, 'what it's like to do so-and-so' only ever means 'what doing so-and-so resembles'; so because there is no way to inform someone what experiences of having a sweet taste resemble, and yet there is still something having a sweet taste is like, 'what having a sweet taste is like' must not be used in its only ordinary way when concatenated to 'know'.[4] I reply that Lewis misses the fact that you can also say what it's like to taste New Zealand beer by saying how the experience is, namely

pleasant. It is incorrect to assume that literal, ordinary uses of 'what it's like to do so-and-so' solely concern what doing so-and-so resembles: they can also concern how doing so-and-so is, what features or properties it has. I expand upon this point below.

The popularity of the non-normality view is somewhat surprising, because, to my knowledge, no one has ever provided or defended a positive theory of how the frame was assigned its non-normal meaning. How might such a theory go? Clearly there is not enough in the frame to hang a metaphor on; nor were we ever given explicit instruction in its meaning, as we tend to be when learning idioms.[5] It seems, then, that such a theory must say that the non-normal meaning is assigned in the manner in which expressions typically acquire technical meaning, that is in something like the following way. Each of us begins with a grasp of the content the frame is used to express. This content becomes highly salient to one when one engages in the philosophy of consciousness. Once this content has become salient, one can establish an internal convention with oneself to use a certain linguistic expression to express this content. If one's peers have undergone a similar process, a tacit convention may arise whereby all attempt to co-ordinate their verbal behavior so as to use the same expression to express this content. Obviously any expression would do, so long as everyone uses it: 'the eagle flies at midnight' would do just as well. However, Nagel 1974 happened to show up at just the right historical moment, and as a result, a tacit convention coagulated in the philosophical community around using this frame to express the relevant content.

Whether uses of the frame in discussions in the philosophy of consciousness are, in fact, non-normal is important, because if they are, it would be very difficult to assess any claim about the logical form of (2). We would no more be able to appeal, in assessing such claims, to facts about the constituent structure of the sentence or the meanings of its parts than we would in assessing the logical form of the code 'the eagle flies at midnight' or the idiom 'John kicked the bucket'. The best we could hope to do would be to appeal to brute intuition concerning the technical meaning, and such appeals have a slim hope of resulting in any sort of consensus.

But fortunately, it's not very plausible to suppose that the frame in its present use acquires its meaning from an implicit technical convention. Obviously the story of how the convention arose requires a great deal of filling in. But the project of filling in this story does not need to be carried out: even these broad outlines are enough for the the story to be empirically refuted in its cradle, because plenty of people who are not plausibly party to the philosophical convention seem to be using the expression in just the way we philosophers are using it. First, the title of Nagel 1974 can be instantly understood by anyone, including those with no exposure to the philosophy of mind. This stands in contrast with the difficulty the uninitiated have understanding the title of Jackson 1982, a contrast which the technical convention view is hard-pressed to explain. Similarly, anyone who has taught undergraduate courses in the philosophy of mind will be aware that, while undergrads have a hard time grasping technical jargon, even such basic pieces of technical jargon as 'substance', 'attribute', and 'supervene', they have no comparable difficulty in understanding (2).

And second, the frame is in wide use in the culture at large, and was so long before the philosophy of mind took on anything like its current shape. Pop songs from the Beatles' 1966 song 'She Said, She Said' ("She said 'I know what it's like to be dead'") to Everlast's 1999 song 'What It's Like' ("God forbid you ever had to walk a mile in his shoes, because then you really might know what it's like to have the blues") employ the frame. So do hundreds of thousands of web pages. Hundreds of web pages even discuss the question of whether a blind person can know or imagine what it is like to see or the desire of a person who went blind to know what it is like to see again. I do not find it particularly plausible that there is any important difference between the phenomenon these pop songs and web pages address and the phenomenon (2) concerns.

* * *

If (2) can be understood through normal means, then it is possible to investigate its meaning in a systematic manner. We can appeal to our tacit grammatical knowledge in recognizing various sentences as transformations of (2), in recognizing the

component expressions of (2) and of these transformations, and in assessing the meanings of these component expressions and how they combine to determine the meaning of the whole.

To begin with, one might wonder, upon looking at (2), "Mary didn't know what *what* is like to see a red thing?" But clearly, 'it' is not intended to refer here: (2) and (2′) 'Mary doesn't know what seeing a red thing is like' are equivalent transformations; all 'it' does in (2) is provide a syntactic subject for 'is' when its semantic subject, the gerund 'seeing a red thing', appears at the end of the sentence in the form of the clause 'to see a red thing'.

Intuitively, 'seeing a red thing' in (2′) concerns experiences of seeing a red thing. Experiences are events. As events in general are particular happenings in the world, such as baseball games or episodes of buttering, so are experiences particular happenings in the world: happenings which are experiencings. Some such events are events of seeing; some events of seeing are events of seeing a red thing. Experiences are events that take place in the mental lives of particular subjects of mentality. An experience is a certain subject's iff the experience takes place in that subject's mental life. In 'seeing a red thing', no subject is specified. Intuitively, it concerns experiences of seeing a red thing in general: normal/typical experiences in normal/typical subjects, within some range. So (2′) is more or less equivalent to 'Mary doesn't know what the typical person's typical experiences of seeing a red thing are like'. This negates (2_{T}):

> **$2_{\mathbf{T}}$** Mary knows what [the typical person's] [typical] experiences of seeing a red thing are like.

I will briefly explain the frame 'what $\cdots$ is like' and the frame '[someone] knows $\cdots$' as applied to a clause that begins with a question-expression. More detailed argumentation concerning these structures is given in an appendix.

'Like what' can be used to ask questions about the properties of things. When it is, this use has nothing to do with the sorts of comparative questions Lewis focused on in the passage discussed above. This seems fairly obvious: when one asks 'what is San Francisco like?' it is appropriate to respond with a string of predicates, such as

'dense, hilly, and expensive'. The comparative question would be more appropriately answered with a string of NPs (Noun Phrases), such as 'uranium, Ithaca, and cuts from the tenderloin'. In certain contexts, this might be an apt answer, but in most, it would be bizarre.

Rather, it seems that, in such cases, 'like what' is used to ask which properties a thing has: it serves as a question expression for predicates, in much the same way as 'who' serves as a question expression for noun phrases denoting people. When I ask 'who let the dogs out?' I want you to provide me with names or descriptions of the person or people who let the dogs out; when I ask 'what is San Francisco like?' I want you to provide me with a predicate or some predicates denoting San Francisco's properties.

It is sometimes argued in the philosophical literature that sentences such as, on the one hand, 'John knows that Italian eggplants are white' or 'John knows that one can get an Italian newspaper at Mayer's', and, on the other, 'John knows what color Italian eggplants are' or 'John knows where one can get an Italian newspaper' must express fundamentally different sorts of facts: that sentences of the former sort involving 'that'-clauses report propositional knowledge, whereas sentences of the latter sort involving *wh*-clauses or "embedded questions" report some other, non-propositional form of knowledge.

Such a claim is not particularly plausible on its face: one can, after all, specify John's knowledge after reporting it with an embedded question: as with 'John knows what color Italian eggplants are—white', or 'John knows where one can get an Italian newspaper—at Mayer's'. It is very hard to see what the purpose of giving such a specification would be if it were not to make determinate the content of John's propositional knowledge which was left undetermined by the initial claim.

Rather, knowledge ascriptions involving embedded questions are intended to ascribe propositional knowledge: knowledge of a proposition which would count, in a context, as an apt answer to the embedded question. So, for instance, the sentence 'John knows where to get an Italian newspaper' communicates that, for some proposition, with a certain contextually salient property—such as concerning

newsstands in Ithaca (knowing that one can get an Italian newspaper in Rome might not count)—which answers the question 'where can one get an Italian newspaper?', John knows that proposition. Such a proposition would have the form ‖one can get an Italian newspaper at NP‖.[6]

So (2_{T}) reports that Mary knows some proposition which answers the question 'what are experiences of seeing a red thing like?'. It is clear that, in the typical context in which one considers the Knowledge Argument, one is concerned with Mary's knowledge of the properties of experiences of seeing a red thing, and not with Mary's knowledge of what things experiences of seeing a red thing are similar to. Hence such a proposition would have the form ‖experiences of seeing a red thing are PRED‖.

Suppose then that (2_{T}) is uttered in a context in which the property of propositions of being F is salient. This utterance would be equivalent to 'there is some truth t such that (i) t is a propositional answer to the question 'which properties do experiences of seeing a red thing have?', (ii) Mary knows t, and (iii) Ft'. This would in turn be equivalent to 'there is some truth t such that (i) t is of the form ‖experiences of seeing a red thing are PRED‖, (ii) Mary knows t, and (iii) Ft'. (2) negates this claim, and is thus equivalent in such a context to 'no truth t is such that (i) t is of the form ‖experiences of seeing a red thing are PRED‖, (ii) t is known by Mary, and (iii) Ft'. This is compatible with Mary's knowing all F truths, if there is no F truth of the relevant form. However, in standard discussions of (2), it seems to be presupposed that there is such a truth, and therefore in context (2) seems to be equivalent to 'there is some F truth of the form ‖experiences of seeing a red thing are PRED‖ which is unknown to Mary'.

This negated claim differs from (2_{lf}) only in that it unpacks the idea of a truth answering a question, and in that the question to be answered, 'what are experiences of seeing a red thing like?', is equivalent to, though distinct from, the question 'what is it like to see a red thing?'.[7]

* * *

Well that's pretty thin. Strip away the jargon, and what I've argued is that (2) says

that there's something about seeing a red thing that Mary didn't know—*you know what I mean.* Nearly all the interesting work of conveying *which* truth about seeing a red thing Mary didn't know needs to be done by context. What then do we mean when we assert (2)?

Although (2) is in itself so insignificant, the fact that we are so drawn to such an insignificant sentence when discussing Mary's situation is itself significant. By refusing to be explicit about the knowledge that Mary lacks, the speaker manages to convey that the knowledge she lacks is of a special sort which can't be put into words. The speaker exploits the listener's ability to carry out the following line of reasoning:

> If you assert of someone that they lack a certain sort of knowledge, but you are inexplicit about what it is, and I know that you know what it is, and I know that you are trying to be as explicit as possible (e.g., you aren't trying to allude to something in a less-than-explicit way because you are playing games, or trying to avoid some bad consequence that would arise from greater explicitness), then I would be reasonable to conclude that you think you can't express this piece of knowledge more explicitly. This might happen because you are having a hard time finding the right words. But usually, in such cases, one hems and haws and makes groping gestures, or apologizes, or talks around the point, or in some other way manifests that one is tongue-tied. So since I don't notice you acting tongue-tied, the only thing to conclude is that you think you aren't alone in your inarticulacy, and that no one can express the piece of knowledge more explicitly: it must be that this piece of knowledge cannot be put into words.

If the listener went through this line of reasoning, this would surely raise to salience, as a feature of the knowledge she lacks, the property of being incapable of being put into words.

This would not stretch the boundaries of common sense: the idea that certain propositions or concepts cannot be put into words, but can only be understood by one who has had a certain experience, is an aspect of our pretheoretic view of the mind.[8] There are a number of places where this pretheoretic idea needs refinement. For instance, there is a sense in which these propositions *can* be put into words: one can say 'this is what it's like to see a red thing!' Such an utterance can

even communicate the intended proposition, if the hearer appropriately sympathizes with the speaker. Perhaps the sense in which these propositions are inexpressible is that they cannot be expressed using words which have robust context-invariant semantic content. And there is a sense in which any (non-innate) concept can only be understood by one who has had an experience of a certain sort, an experience which suffices for learning that concept: arguably, in order to grasp the concept 'bachelor', one needs to have an experience of having it defined for one, or perhaps an experience of observing the ostension of a number of paradigms and foils. However, there's no room to provide this refinement here, let alone to address the (interesting and important) question of just what distinguishes grasp of expressible from grasp of inexpressible concepts. I do think that the idea is familiar and serviceable as pretheoretically understood, so I will work with it.

One important class of inexpressible propositions is involved with reflection on our own mental states and sympathetic understanding of those of others.[9] Human adults have the ability to simulate in themselves experiences they are not actually having, such as when one imagines oneself riding a rollercoaster. These experiences do not have the same connections to actual belief and action that genuine experiences have, but they are useful in establishing hypothetical or suppositional lines of thought. This ability to simulate experiences tends to be somewhat limited, in that for the most part one can only simulate an experience of a certain type if one has actually had an experience of a similar type. Human adults also have a general ability to reflect on the character of their own experiences,[10] or of experiences they simulate in themselves, where reflecting in this way brings with it a distinctive way of understanding the character of the experience unavailable to one who has not had or simulated an experience with that character. Once one has this sort of reflective understanding of an experience with a certain character, but not otherwise, one can also sympathetically understand that character as a feature of the experiences of others.[11]

Novel episodes of such understanding seem to be able to underlie the entertaining of thoughts with novel propositional content; for one thing, they seem to have

distinctive powers to explain other facts that explanations "in words" lack; and it seems plausible that what explains are propositions (Williamson, 2000, §9.5). Clear cases of this novel explanatory power emerge in cases in which one is unable to understand why people with certain experiences feel or behave a certain way: for instance, why many people who are very rich are not happy, why many people who live in areas with above-average rates of street crime do not support more aggressive policing techniques, or why why anyone would vote for a certain presidential candidate. One might be presented with certain fairly obvious explanations: it might be explained that a feeling of overcoming challenge or adversity through hard work is rewarding; or that up to a certain point the rate of street crime can increase without demanding much more than slightly increased vigilance, whereas aggressive behavior by arrogant police officers from outside the neighborhood can damage one's sense of pride in community and subject law-abiding residents to the risk of being demeaned by the police; or that their voting behavior is solely determined by the candidate's emotional appeals because they are incapable of understanding the deleterious impact his policies are likely to have. One might hear these explanations, but still feel as though the behavior does not make sense. However, if one were to have or simulate the person's experience—for instance, by taking a month off work to go on a pleasure cruise, or by being treated rudely by a policeman at a traffic stop, or by recalling one's juvenile feelings of patriotism and admiration of power—the behavior might suddenly make sense to one.

This all seems also to be an aspect of our pretheoretic view of mind. Once again, there are nice questions here, which there is no room to address: such as just what grasp of reflective/sympathetic concepts consists in such that thoughts containing them yield explanations that expressible concepts cannot.[12] Still, the presence of this additional explanatory power provides strong support for the view that reflective/sympathetic concepts are genuine concepts, and can enter distinctively into thoughts with propositional content.

I thus propose that what is conveyed by (2) (in the present context) is that, concerning a certain reflective/sympathetic proposition about the nature of experi-

ences of seeing a red thing, Mary didn't know it. Call this proposition "𝔱". Which proposition is 𝔱? Supposing that Mary is a normal subject, and hence much like you, you can entertain 𝔱, or at least a proposition much like it,[13] through the following procedure. Cause or simulate in yourself an experience of seeing a red thing, and then focus your attention on the character of the experience.[14] Think to yourself: seeing a red thing is like this.

* * *

The link between inexpressible reflective/sympathetic propositions and "knowing what it's like" has provided much inspiration to writers of pop lyrics, a great many of which complain that someone doesn't know what a certain sort of experience is like. Standardly the singer has been jilted, and intends to convey that his or her emotional distress is extraordinarily intense. Such lyrics depend for getting their point across on one's ability to reason as follows:

> There's some proposition about the character of the singer's experience of emotional distress I don't know; but since the singer won't tell me exactly what it is, it must be inexpressible; so it must be a reflective/sympathetic proposition; the singer seems to be lamenting my inability to know it; but since the singer won't do anything else to convey it to me, the singer must think I can't grasp it; if I can't grasp it, I haven't had an experience of emotional distress that substantially resembles the singer's experience; the best explanation for this is that the singer's emotional distress is extraordinarily intense.[15]

It is a virtue of my analysis of (2) that it so readily explains the popularity of this lyrical trope.

3 Assessing the Knowledge Argument

The following points are supported by the preceding discussion. I located a certain class of concepts, the reflective/sympathetic concepts. A reflective/sympathetic concept C_X is a predicate, concerning the character X of experiences of a certain type T_X. Reflective/sympathetic concepts are, in some intuitive sense, "inexpressible": this is to say that in general, and subject to the qualifications detailed above, one

cannot grasp C_X unless one has had an experience of type T_X. There is a type T_R of experiences of seeing a red thing, which have a character R, and a corresponding reflective/sympathetic concept C_R. Before she is released, Mary has never had an experience of type T_R; consequently she cannot grasp the concept C_R. Standardly, when one says that someone does or does not know what it is like to see a red thing, one means that the person does or does not know the proposition $\mathfrak{t}$ predicating C_R of experiences of type T_R. Because Mary does not grasp the concept C_R, she cannot entertain any proposition involving C_R, so *a fortiori* she cannot know a proposition predicating C_R of experiences of type T_R: hence she can't know what it is like to see a red thing.

The Knowledge Argument can then be recast in a somewhat simpler form as arguing validly from the premisses that (1) for every physical truth, Mary knows it, and (2′) Mary does not know the truth $\mathfrak{t}$, to the conclusion that (C′) $\mathfrak{t}$ is a non-physical truth.[16] (C′) is supposed to suffice for dualism. If a non-physical truth is a truth which concerns the condition of entities other than physical things, it plausibly does, given the form of $\mathfrak{t}$.

How successful is the Knowledge Argument as an argument for dualism? Any attempt to answer this question is subject, on pain of falling into either anti-naturalism or skepticism, to a requirement of psychological plausibility: it must capture how we think when we think through the Knowledge Argument. Of course a central aspect of our reaction to the Knowledge Argument is that it strikes us, pretheoretically, as successful—or, at least, if it is unsuccessful, it is far from obvious wherein its lack of success lies: Stoljar and Nagasawa (2003) list five distinct lines of reply to the Knowledge Argument, each with its staunch adherents; obviously, the physicalist opposition is mired in internecine strife—not an easy position from which to make accusations of an obvious fallacy! So any assessment of the Knowledge Argument according to which it doesn't involve some hard-to-unearth error is automatically cast into doubt.

* * *

Stoljar and Nagasawa list five major physicalist accusations against the Knowledge

Argument. The first denies (2′); the fifth denies (1). Denying (2′) strikes me as pointless for reasons that should become more clear below; I discuss (1) below. The remaining three allege that the argument is invalid as an argument for dualism due to some fallacy of equivocation: we read the premisses in one way and find them plausible; we read the conclusion in a different way and find it interesting.

Diagnoses due to equivocation generally face a problem of asymmetry. If the premisses are plausible on disambiguation d_1 but not d_2 and the conclusion is interesting on disambiguation d_2 but not d_1, why aren't we equally likely to find the argument for the interesting conclusion implausible, because rather than reading the argument with d_1 premisses and d_2 conclusion, we have read it with d_2, d_2; or plausible but uninteresting, because we have read it with d_1, d_1; or implausible and uninteresting, because we have read it with d_2, d_1?

Two of the accusations of equivocation concern an alleged equivocation in 'know'. Even if 'know' is not univocal between uses in which it takes a clause and uses in which it takes a direct object, it is certainly univocal in uses in which it takes a clause.[17]

The remaining accusation of equivocation (which is probably the most popular physicalist reply to the Knowledge Argument) appeals to an alleged distinction between a "Fregean" and a "Russellian" conception of truths, where this distinction amounts to whether, given some entities "bundled together in a worldly state of affairs", there can (Fregean) or cannot (Russellian) be more than one truth concerning this bundle. (1) and (2′) are plausibly psychological when read along "Fregean" lines; (C) is interestingly ontological when read along "Russellian" lines. Allegedly, we succumb, in considering the Knowledge Argument, to a tendency to equivocate between these conceptions of truths.

But first, it's hard to imagine that this tendency, if it in fact exists, would have been allowed to slip by: the distinction between Fregeanism and Russellianism is incredibly familiar in philosophy. Worse still, as can be seen in the passages I displayed in §1, Jackson was explicitly aware of opacity concerns in the 1986 paper. It would take a major helping of doublethink to first explicitly cleave to Fregeanism

in the statement of the conclusion and then reject it in drawing out the significance of the conclusion.

Second, I doubt that this tendency exists. Although some philosophers advocate that we sometimes or always think of truths as Russellian, I disagree. I never find myself unconsciously slipping into accepting that Lois knows the truth that Clark flies around Metropolis. It is hard to understand what could be the difference between my situation with respect to this truth and the situation of those who regard the Knowledge Argument as an argument for dualism, such that this difference could ground the alleged distinction in psychological outcomes of these situations.

* * *

In recent work, Jackson (1998) has come up with a way of extracting dualism from the Knowledge Argument which purports to reveal what the argument was after all along. Stoljar and Nagasawa (2003, §3.4) boil down this approach as follows: (i) "if physicalism is true, the psychophysical conditional [‖if P then Q‖, where P "gather[s] together all the [obviously] physical truths of the world into one mega-truth" and Q gathers together all the reflective/sympathetic truths of the world into one mega-truth] is *a priori*" and (ii) there are no *a priori* connections between obviously physical and reflective/sympathetic propositions.

Whether or not these claims are true, their relation to the Knowledge Argument as I presented it, and to Jackson's early presentations, is not obvious. Some further explanation of how, pretheoretically, we manage to deduce (i) and (ii) from the principles of the Knowledge Argument, or regard them as licensed by the Black-and-White Mary Scenario, is required to make this stick as an explanation of the allure of the Knowledge Argument.

Moreover, it is not clear that claim (ii) is either true or supported by the Black-and-White Mary Scenario. It is widely recognized that before her release, no amount of calculation would have enabled Mary to know what it's like to see a red thing. But her problem was not that, like a frustrated mathematician trying to prove Goldbach's conjecture, she was able to entertain $\mathfrak{t}$, but couldn't prove it. Unlike the frustrated mathematician—who may spend every waking hour wondering about

Goldbach's conjecture—Mary wasn't even in a position to wonder whether t is true. Conjure in yourself an experience of seeing a red thing, focus on its character, and think to yourself 'I wonder whether this is what it's like to see a red thing'. Mary couldn't have done that, before her release. Clearly at least one source of her problem with knowing t was that she couldn't take the constitutive step of even entertaining t. The question whether t is *a priori* entailed by obviously physical propositions cannot be tested by considering someone who can't even entertain t: a case could only count as supporting the negative answer if it concerned someone who could entertain t, and who found himself incapable of demonstrating it by *a priori* reasoning from his obviously physical knowledge.

Wait until Mary goes to sleep some night, then perform super-neurosurgery on her so that she gains the ability to visualize a red thing. This would give her all the concepts required for entertaining t; but would this alone suffice for her coming to know that t is true—as opposed, for instance, to its remaining open for her whether this is what it's like to see a green thing? Aside from the significant departure this example represents from the original Black-and-White Mary Scenario, I find that my intuitions here are dim—though if anything, it seems she could rule out that she was imagining a green thing on the basis of her knowledge that green things look "cool" and that the thing she imagines would look "hot" (Hardin, 1997). The dialectic that emerges from this point is complex: see Hilbert and Kalderon 2000 for some of the ins and outs. It seems to me that the question remains open, and extremely vexing, whether there are some obviously physical propositions which *a priori* entail t.

* * *

If the Knowledge Argument is valid, and (2′) is (as I take it to be) undeniable, then dualism follows if (1) is true. A committed physicalist has an easy out: she can argue from (2′) and the denial of (C′) to the denial of (1). The denial of (1) in the face of the Black-and-White Mary Scenario can be supported by the familiar "Fregean" observation that several distinct truths can concern the same "worldly state of affairs": it can be claimed that there is some physical truth Mary already

knew which concerns the same condition as $\mathfrak{t}$ concerns. $\mathfrak{t}$ is a physical truth; Mary could not have known all the physical truths before her release.

But one who says this is at risk of saying something highly unintuitive. The Knowledge Argument is, after all, alluring, and the view that Mary can learn all the physical truths before her release is a particularly alluring subcomponent. None of my undergraduate students in Intro Mind have ever objected to this step. Even "post-theoretic" subjects ignore this option: of the five distinct classes of physicalist reply to the Knowledge Argument detailed by Stoljar and Nagasawa (2003), all but the last ignore this reply (and a tiny minority of the articles cited here address this point). Philosophers seem to regard this as a last-ditch escape from dualism, to be adopted only if all other options are exhausted.

There are two reasons why this reply might lack appeal: first, it might not occur to us, because we have some mental block against it; second, we might find it implausible even if it occurs to us. On the first option: we might somehow be distracted from wondering whether $\mathfrak{t}$ is a physical truth. An intriguing possibility for the source of this distraction (suggested to me by Zoltán Szabó) is that this distraction results from the fact that what we are most strongly attending to in assessing the Knowledge Argument is what Mary *knows,* rather than what is *physical.* The claim that there is such an imbalance in attention is plausible: it's the *Knowledge* Argument; the article is titled 'What Mary Didn't *Know*'; etc. As a result of this imbalance in attention, the questions we will tend to ask about the Knowledge Argument will tend to concern the scope or nature of Mary's knowledge, rather than the scope or nature of being physical. Stoljar and Nagasawa's survey makes this plausible: four of the five physicalist replies, comprising more than four-fifths of the cited articles, concern the scope or nature of Mary's knowledge.

On the second option: we might simply find it less plausible that $\mathfrak{t}$ is physical than that (1) is true.[18] Presumably this is because there is some property such that we regard the possession of that property by a proposition as a necessary condition for that proposition's being physical, a property which does not hold of any reflective/sympathetic proposition (note that nothing is special about $\mathfrak{t}$: any

reflective/sympathetic proposition would do). It is hard to see what property this could be, aside from the property of being expressible: we know little or nothing of the specific content of Mary's lessons, nor do we care to know anything in particular about it. As Lewis (1988, p. 281) notes, "[l]essons on the aura of Vegemite will do no more for us than lessons on its chemical composition. [...] Our intuitive starting point wasn't just that *physics* lessons couldn't help the inexperienced to know what it's like. It was that *lessons* couldn't help". One thing can be certain about this content, however: because it is conveyed in lessons, it is expressible.[19]

The empiricist thesis that an expressible concept and an inexpressible concept cannot both denote the same entity has a long tradition in the philosophical literature, tracing back, perhaps, at least to Hume (1739/1978, I.I.i).[20] If this is the core idea behind the Knowledge Argument, the passage through knowledge is largely a detour: a more efficient way to express the core of the argument would be to appeal to this empiricist thesis, together with the thesis that, for any physical entity, there is an expressible concept denoting it.

Assessing the empiricist thesis is the work of another paper, or perhaps a long book; my primary aim in this paper has been to tease out its hidden influence on our reaction to the Knowledge Argument. However, I will make some (somewhat speculative) evaluative remarks here. A recurrent idea in the philosophical literature is that certain predicate concepts "reveal the natures" of the properties to which they refer (Johnston, 1992, pp. 138–9; Lewis, 1995, pp. 327–8; Chalmers, 2002, p. 256). Contrast such concepts as 'water': one can grasp this concept in complete ignorance of chemistry; and yet, it seems that chemistry reveals the nature of water when it tells us that to be water is to be (or to be constituted by) H_2O. Grasp of a nature-revealing concept, on the other hand, allegedly does not allow for such ignorance about the nature of its referent: such grasp allows only for ignorance about such "extrinsic" matters as its pattern of instantiation.

In this literature, reflective/sympathetic concepts are frequently linked with this idea of Revelation. One who takes a reflective/sympathetic concept to be revelatory in this way might naturally accept that it cannot co-denote with an expressible

concept. For suppose that C_x is an expressible concept. Perhaps, possession of a given reflective/sympathetic concept C_r is compatible with ignorance of ‖to be C_r is to be C_x‖, whatever C_x may be.[21] If so, then, given Revelation, that proposition is false.

Why do we feel that there is a link between inexpressibility and Revelation? This opinion can be explained in at least two distinct ways. The first goes by reflection on an allegedly standard pattern of discovery of natures.[22] The concept 'water' allegedly encodes causal structure: water is that substance which plays a certain causal role. And the concept 'H_2O' allegedly encodes causal structure: H_2O is a substance which, as a matter of fact, realizes that causal role. No discovery of nature can be in any way licensed without beginning with a concept which encodes causal structure: without this, the reducing concept cannot get a "toehold". But if a concept is inexpressible, it does not encode causal structure: for causal structure is always expressible.

The second goes by reflection on the functional role of an reflective/sympathetic concept. Perhaps the referent of the concept is somehow "incorporated into" the concept itself: perhaps deploying the concept in a thought involves running a computational process of a certain type, some subprocess of which is of a type which is the referent of the concept; perhaps grasp of the concept requires a primitive relation of "acquaintance" with the referent. Such a concept would be inexpressible, since either way grasp of the concept would essentially involve having an experience of a certain type. And, moreover, such a concept might contribute a feeling of revelation: since the referent would be, in a sense, part of the concept, the concept would not "stand between" the subject and the referent; rather, the referent would be "present to the mind" of the subject. On this view, incorporation of the referent by an reflective/sympathetic concept lies at the root of both the inexpressibility of the concept and the feeling that it provides Revelation.

* * *

Suppose that it is more *prima facie* plausible that Mary could learn all the physical truths before her release than it is that t is physical, perhaps because we are in fact

deeply attracted to the empiricist thesis. Presumably most physicalist philosophers are physicalists because they take it that nothing non-physical can be involved in the causal order; and because they take it that the characters of experiences are involved in the causal order. If this argument is compelling, we would then be faced with an antinomy: the characters of experiences must be physical, because they are involved in the causal order; but they cannot be physical, because they can be denoted with inexpressible concepts. This is a serious problem, and I don't see how to make it go away: in part this is because the empiricist thesis has not been adequately investigated. I hope to have shown, however, that this investigation is required. For grant the empiricist thesis, and physicalism falls: it won't do to go out in search of some "fallacy", because there is no fallacy in the neighborhood. One can, of course, stabilize one's epistemic state by turning a blind eye on the antinomy and pleading "antecedent physicalism". But, I hope, most philosophers will find this option uncongenial.

Appendix: Further Formal Properties of (2)

In this appendix, I argue in greater detail for my claims about the formal, logical properties of (2).

The credibility of my claim that 'like what' can be used to ask about the properties of things can be given an additional boost by deriving this behavior from a theory of the deep linguistic properties of 'like what'. I begin with an analysis of the related construction 'like that'.

The familiar "comparative" use of 'like this/that', meaning 'similar to this/that', is composed out of 'like' meaning 'similar to' and the singular pronoun 'this/that'.[23]

But alongside this comparative use, there is also a "pro-predicative" use of 'like this/that'. A "pro-predicate" or "predicative proform" is an expression like a pronoun in lacking intrinsic semantic content—in the sense that 'he', 'it', and 'then' have no substantial fixed meaning or denotation, unlike 'dog', 'run', or 'and'—and thereby needing to inherit it from outside (e.g., by demonstration, anaphora, or

binding), but which appears in predicate rather than in argument position. 'Like this' seems to have taken over the role as preferred English pro-predicate from the now somewhat archaic 'thus'. I shall defend three claims about pro-predicate uses: (I) 'like' does not mean 'similar to'; (II) 'this/that' is not used as a singular pronoun; and (III) 'like this' is semantically incomposite; it follows that the comparative and pro-predicate uses are totally unrelated.

Consider a demonstrative use of pro-predicate 'like this': demonstrating red, one utters (d) 'my couch is colored like this'. Here, 'like this' is not used comparatively, but is rather used as a semantically incomposite demonstrative pro-predicate, referring to red in virtue of one's demonstration.[24] Concerning (I): first, it is intuitively clear that a comparison between the couch and the bearer of the demonstrated instance of redness need not be intended here, and that the utterance may just mean the same as the non-comparative 'my couch is colored red'; so, intuitively, 'like' does not mean 'similar to'. Of course 'my couch is like this' and 'my couch is similar to this object' (said, in the first case, focusing on a property-instance and, in the second case, focusing on its bearer) are necessarily equivalent and obviously so, so it may be easy to convey what is said by uttering either by uttering the other. But necessary equivalence is not, as is familiar, equivalence of meaning. And, according to my intuitive sense of the meaning of the utterances, the former must convey that my couch has the designated property and need not convey that it is similar to any particular object, whereas the latter must convey that my couch is similar to a certain object in a certain respect, but need not convey which property they have in common.[25] Second, 'colored' seems to want to be followed by an adjective, as in 'my couch is colored red'/*'my couch is colored Fred', so that explicit comparatives do not sit happily next to it: *'my couch is colored resembling that'/?'my couch is colored similar to that'; by contrast, (d) does not suffer from this problem.

Concerning (II): first, the process of understanding an utterance of (d) differs from the process of understanding a comparative use. In order to understand the utterance of (d), one's audience must merely determine which color one is talking about. By contrast, in order to understand an utterance of (for instance) 'my couch

is similar in color to that material body', one's audience must (i) determine which material body in the region of ostension one is talking about, and (ii) determine what its color is. Second, one can wave at a whole range of red objects and sensibly utter (d). By contrast, if one waved at a range of red objects, it would be bizarre to utter %'my couch is similar in color to that'.[26] One's audience might be able to understand the claim, since all the waved-at objects are the same color. But the pronoun could not refer to the color: *'my couch is similar in color to red/redness' is unacceptable. And it could not refer to any individual, since no individual is designated. So it could not refer at all. Only a grammatically plural pronoun could refer given this sort of gesture, as in 'my couch is similar in color to those'.

Concerning (III): one might agree that 'that' is not a pronoun but a pro-predicate used to denote red, but dispute my claim that 'like that' is a semantically incomposite pro-predicate, thinking that 'like' has some other function. But 'like' must disappear along with 'that' when the proform is cashed out: *'my couch is like red', unlike 'my couch is, like, red', is unacceptable.

Next, consider a discourse-anaphoric use of pro-predicate 'like that', in the discourse (a) 'the couches here are red'; 'couches which are colored like that please me'. Here, 'like that' is not used comparatively, but is rather used as a semantically incomposite anaphoric pro-predicate, picking up its meaning from the use of 'red' earlier in the discourse. Concerning (I): here, it is also intuitively clear that no comparison need be intended. Concerning (II): the first and second arguments above carry over trivially to this case. An analogue to the third argument (from plurals) runs as follows: in (a), the only options for the antecedent to the anaphor are the plural NP 'the couches' and the predicate 'red'. But a plural NP cannot be the antecedent of a grammatically singular discourse-anaphoric pronoun: 'the people here are your friends'; '(they buy/*he buys) me a drink regularly'. So the antecedent of the anaphor in (a) can't be the NP, but must be the predicate; hence the anaphoric proform is not a pronoun but a pro-predicate.[27] Concerning (III): arguing as above.

Finally, consider a bound use of pro-predicate 'like that'. The example in this

case is somewhat harder to hear in the intended way, so I'll set it up a bit more extensively. Suppose I tell you that I have a rug with squares in Southwestern colors. The following dialogue might ensue: 'Does it have a square which is colored adobe tan?'; 'Yes, my rug has a square which is colored like that'; 'Does it have a square which is colored sage green?'; 'Yes, my rug has a square which is colored like that'; 'Does it have a square which is colored arid-sky blue?'; 'Yes, my rug has a square which is colored like that'; 'Does it have a square which is colored cactus-flower red?; 'Yes, yes—[(b)] for *every* Southwestern color, my rug has a square which is colored like that'.[28] Here, 'like that' is not used comparatively, but is rather used as a semantically incomposite bound pro-predicate, ranging over the Southwestern colors in the domain of the NP 'every Southwestern color'. Concerning (I): here we do not need to appeal to intuition to note that no comparison is intended, for interpreting (b) with 'like' as 'similar to' and 'that' as ranging over Southwestern colors results in mumbo-jumbo. Squares are parts of my rug, concrete areas of tufting; colors are (if anything) universals. It would be perverse to compare an area of tufting with a universal, which is why %'for every Southwestern color, my rug has a square which is similar to that' is semantically dreadful. Since (b) is not semantically dreadful, clearly no comparison to the entities in the domain of the NP is intended. Concerning (II): the entities in the domain of the NP are properties, and properties are given either with predicates or with such nominalizations as 'redness'. But in the latter case, the matrix clause would not have a sensible predicate: *'my rug has a square which is (colored) (like) redness' is unacceptable; hence, the bound proform must be a pro-predicate. Concerning (III): arguing as above.

Proforms are closely connected with such "question words" or "*wh* forms" as 'who', 'what', and 'where'. Each proform is paired with a *wh* form. In English, the members of this pair tend to sound similar (Lycan (1995, p. 245) presents as examples of this phenomenon the pairs 'then'/'when', 'there'/'where', and 'thither'/'whither'; compare also 'what'/'that'),[29] and, in the most fundamental sense recognized by standard contemporary views in syntax,[30] can occur in the same syntactic positions ('I boiled the potato then'/'when did I boil the potato?'; 'I put the potato

there'/'where did I put the potato?'; 'I boiled that'/'what did I boil?'). A proform and the *wh* form with which it is paired both lack intrinsic semantic content, but instead range arbitrarily over a certain class of entities, with the members of the pair ranging over the same class: 'where' and 'there' both range over places; 'when' and 'then' both range over times or conditions; 'whither' and 'thither' both range over paths; and 'what' and 'that' both range over arguments. The proform and the *wh* form differ only in that the proform has some entity from the range assigned to it by linguistic or extralinguistic context, whereas the *wh* form does not have an entity assigned to it but is rather used in asking questions about that range of entities. An appropriate answer to a question consists of a word or string of words denoting entities chosen from this range ('when did you boil the potato?' —'from 3PM to 4PM'; 'where did you put the potatoes?' —'in the lower cabinet, in the basket, and in the boiling water'; 'what did you boil?' —'the potato, the radish, and the cabbage').

We should thus predict as an instance of this generalization that there would be a *wh* predicate paired with pro-predicate 'like that'.[31] The *wh* predicate would be pronounced 'like what', would appear syntactically in predicate position, and would be used to ask questions about predicates. This prediction is confirmed: alongside the comparative use, 'like what' is also used as a predicate *wh* form. Just as one can say 'San Francisco is like that' with pro-predicate 'like that', one can ask 'What is San Francisco like?' with *wh* predicate 'like what'.[32] When one does so, an appropriate answer would consist of a predicate or string of predicates such as 'dense, hilly, and expensive'. And, as I noted at the outset, this would be an appropriate answer to this question.

* * *

Stanley and Williamson (2001) have recently argued convincingly for the claim that, when an embedded question appears as the complement of 'know', what is reported is propositional knowledge. They argue that this claim is supported by robust results in cognitive syntax and semantics. Here is a brief overview of these results.[33]

Every embedded question corresponds[34] to an "answer-set" of propositions. Which

propositions are in this answer-set is related to the syntactic structure of the question in the following way. One can think of the question as corresponding to a string that is like a standard sentence, but which has the question-word standing in for some other word or words (e.g., 'one can get an Italian newspaper at Mayer's'/'one can get an Italian newspaper where'). And then one can think of the question word as a variable: the resulting string is then an open sentence (e.g., 'one can get an Italian newspaper x'). This open sentence does not express a proposition or have a complete sense, since variables do not have senses. But it does have an incomplete sense, and this incomplete sense together with how the world is determines which propositions are in the answer-set: a proposition is in the answer-set iff it is a true completion of this incomplete sense.

So, for instance, the answer-set for the embedded question 'what color Italian eggplants are' has as its only member the proposition that Italian eggplants are white; the answer-set for 'where one can get an Italian newspaper' contains the proposition that one can get an Italian newspaper at Mayer's, the proposition that one can get an Italian newspaper in Rome, the proposition that one can get an Italian newspaper on Arthur Ave., and so on. (The same goes for the notion of the set of answers to a "root question", which is the sort of sentence uttered in an interrogative speech-act, such as 'what color are Italian eggplants?' and 'where can one get an Italian newspaper?')

Then, the truth-conditions for an utterance, in a context, of a knowledge-ascription involving an embedded question with answer-set A are as follows. Such an utterance is true iff the subject of the knowledge-ascription knows either some, or all, of those members of A which have a property raised to salience in the context (where whether some or all is required varies from context to context). The contextually variable salience restriction is required. Even if John knows that one can get an Italian newspaper in Rome, 'John knows where to get an Italian newspaper' may be false in a context in which only those members of the answer-set are of interest which concern where to get an Italian newspaper around here. The claim that the members of the answer-set are *propositions* is also required. Even if Lex Luthor

knows that Clark Kent works at the Daily Planet, 'Lex knows who works at the Daily Planet' may be false in a context in which only those members of the answer-set are of interest which concern the day-jobs of participants in the Justice League of America. If all this is correct, the view that 'knowledge-what' is not propositional is in error.

* * *

Here's a further argument for the truth-conditions I've assigned to (2). Consider a more ordinary case in which one is asking someone, concerning a certain unfamiliar experience—climbing a mountain, or being in a riot, for instance—'what was it like?' Standardly, an appropriate answer would be something like 'exhilarating'/'terrifying'/'breathtaking'/'enlightening'. (The significance of this phenomenon is highlighted in Lormand in preparation.) On the other hand, it would be somewhat peculiar if one's interlocutor replied with an NP, such as 'the Tour de France'/'1968'. This is because standardly, when one asks this question, one wishes for a characterization of the features of the experience, of the sort that would be given by a predicate. However, if what was desired were a comparison, it would only be possible to give an NP answer, and a predicate answer would be very bad. Compare: 'what/which cat is your cat like?' 'Otto'/*'exhilarating'. Given that predicate answers are preferable, let alone permissible, I conclude that in these ordinary cases, 'like what' is not used with its comparative reading. I can of course imagine finding an NP answer appropriate. But I think that I would hear it as elliptical for the comparative predicates 'like the Tour de France'/'like 1968'. What is important here is not that noun phrases cannot be given as answers, but that predicates can be given. In a case in which a comparison is explicitly desired, predicates cannot be given. And I cannot detect any difference between these ordinary cases and Mary's case.

Notes

[1]Concerning, of course, superstudent Mary, who, though confined to a room where everything is black, white, or grey, nonetheless acquires a total knowledge of a completed physical (throughout, one may add 'and/or functional', or substitute 'structural' or 'dispositional', making changes as appropriate, without altering the truth of anything I say) science of color and color perception; who is freed, to finally set eyes upon a red thing; who thereupon learns what it is like to see a red thing; and who therefore did not previously know what it is like to see a red thing.

[2]By saying that so-and-so "represents the logical form of" some expression, I mean that so-and-so conveys the meaning of the expression in a way that foregrounds the expression's broadly logical properties and subordinates other aspects of its meaning.

[3]Although all I need to make my point is that this is *a* legitimate way of using 'truth', I also accept the stronger claim that there is no legitimate way of using the ordinary language expression 'truth' which violates the principle that truths are the objects of knowledge. Sometimes philosophers use 'truth' (or, more frequently, 'fact') in such a way that truths are truth-makers for true propositions. But it seems to me that if common sense ontology recognizes entities whose pattern of existence and non-existence determines which propositions are true and which false, these entities would be something like states or events, such as my couch's redness or the World Series. And it seems to be a linguistic error to use 'truth' to apply to states or events: 'Joan already knew all the truths/facts discussed in the chemistry class'/*'Joan already knew all the states/events discussed in the chemistry class'; 'my couch's redness is a pleasing state'/*'my couch's redness is a pleasing truth/fact'; 'the World Series is a thrilling event'/*'the World Series is a thrilling truth/fact'. (The asterisks represent syntactic anomalousness.)

If this is correct, the claim often made by advocates of the opacity line that Mary learned an old fact in a new way is incoherent. Since 'x learned that $p$at t'

implies 'x did not know that p immediately before t', the claim that she learned a fact she already knew implies the claim that she didn't and did know some fact immediately before t. I may be misinterpreting advocates of this view, however; they may be intending to give up ordinary language and common-sense ontology for some superior replacements.

[4]Thanks to Mike Fara and Kieran Setiya for helping me to figure out Lewis's argument.

[5]It will turn out that the phrase gets its significance due to a considerable amount of reasoning about speaker intentions, but this reasoning proceeds on the basis of a prior understanding of the phrase through normal linguistic processes.

[6]Doubled verticals represent "sense–quasi-quotation", so that $\ulcorner\|\varepsilon\|\urcorner$ represents the sense of ε.

[7]Lycan (1995) also appeals to the approach to embedded questions outlined here in the analysis of (2). In Lycan's view, "'S knows what it's like to see blue' means roughly 'S knows that it is like Q to see blue', where 'Q' suitably names some inner phenomenal property or condition" (p. 245). This is unclear: what kind of expression names a property? An ordinary proper name, like 'Homer'? Or a definite description denoting a property, like 'the property of redness'? Or a nominalized adjective, like 'redness'? None of these alternatives yields a grammatical sentence when substituted in for 'Q' in Lycan's frame: *'S knows that it is like Homer to see blue'; *'S knows that it is like the property redness to see blue'; *'S knows that it is like redness to see blue'. The problem is that that 'like' is part of the *wh* form, so that if the answer is given without appeal to a proform (like 'that'), 'like' must go away. So Lycan should say "'S knows what it's like to see blue' means roughly 'S knows that it is Q to see blue', where 'Q' is a suitable predicate denoting an inner phenomenal property". But even still, Lycan's proposal is in one way unmotivated and in another unclear: (i) Lycan does not make it clear what mechanism is at work in ensuring that 'Q' denotes an inner phenomenal property, and (ii) the notion of

"suitability" is left unexplained.

[8]Compare Lewis 1988, p. 262: "They say that experience is the best teacher, and the classroom is no substitute for Real Life". Lewis of course does not himself think that such knowledge is propositional or conceptual, though the platitude he mentions amounts to my claim when interpreted in the most natural way.

Daniel Stoljar points out that Swampman forces this claim to be qualified: compare Lewis 1988, pp. 264–5. To handle such concerns, I could add 'or who is from the standpoint of narrowly individuated apparent memory indistinguishable from one who has had the experience', or make explicit that 'can' is tacitly restricted to normal contexts in this platitude, and that outside of normal contexts, it is unclear what should be said.

[9]Stanley and Williamson (2001, pp. 428–30) discuss another class of inexpressible propositions, those one knows when one knows how to do something. Lewis's platitude applies just as well to these. However these are less plausible as relevant in the present context than are reflective/sympathetic propositions.

[10]By "the character of an experience" I'll always intend to pick out the so-called "phenomenal" character of the experience, that feature of experience which is revealed to reflective attention.

[11]Harman (1996, p. 259) puts the point nicely: "there is a distinctive kind of understanding that consists in finding an equivalent in one's own case. That is "knowing what it is like to have that experience" ".

[12]A substantial obstacle to progress on these questions has been the opinion that reflective/sympathetic concepts are "indexical" concepts. The notion of indexicality was initially presented as a theoretical notion in theories of natural language semantics intended to explain how I assign truth-conditions to your utterances wrt a context. What relevance such a theory of interpretation is supposed to have to the project of explaining the nature of concepts is, I find, exceedingly unclear; as a

result it is unclear how to retrofit this notion of indexicality to the purposes of this project.

[13]It might be impossible for you to entertain $\mathfrak{t}$ exactly. The character of Mary's experience may differ subtly from the character of your experience. Mary's macula may be somewhat more or less yellowed with age than yours, for instance, causing her to regard certain objects you regard as purplish-red as orangey-red, or vice versa. Or red things may be more salient to Mary than to you, since she waited all these years to see one. And your reflective/sympathetic concept 'like this' may be constitutively linked to the character of your experience; so that the proposition she entertains when she thinks 'seeing a red thing is like this' may differ from the proposition you thereby entertain. But, for the most part, we ignore such subtle interpersonal differences in sympathetic understanding. Typically somewhat crude and imprecise understanding meets our purposes in sympathy.

If Mary is not a normal subject, but is rather spectrally inverted, for instance, she does not come to know what it's like to see a red thing: this knowledge concerns what the typical experience of a typical subject is like upon seeing a red thing. If she were aware of her inversion, she could come to know what it is like to see a green thing. In any event, she could come to know what it is like for *her* to see a red thing. Since accounting for this complication would do nothing but encumber the discussion, henceforth I'll ignore it.

[14]There's a considerable simplification here: the experience has a great many characters. What I want you to focus on is the character that experience has solely in virtue of being an experience of a red thing, the character that experience has in common with all other experiences, in normal subjects, of seeing a red thing. This character will be highly abstract and "determinable"; the character of my present experience of seeing a fire-engine red floppy disk under dim light is a "determinate" of this character. Since accounting for this complication would also do nothing but encumber the discussion, I'll ignore it as well.

[15]A less plausible alternative is that the emotional distress is weird: perhaps of the sort one experiences when one's pet bat eats one's pet beetle.

Zoltán Szabó calls my attention to the Patsy Cline lyric 'You don't know the meaning of the words 'I love you so' until you find your true love—and then you know'. He suggests that 'so' here functions as a pro-adverb, specifying a way of loving using, in context, a reflective/sympathetic concept.

[16]This is not fully adequate: here's a better way. Say that a truth is "simple" iff it is either atomic or results from quantifying some or all of the argument places of an atomic truth with quantifiers of positive polarity. The Knowledge Argument can then be recast in a somewhat simpler form as arguing validly from the premisses that (1) for every physical truth, Mary knows it, and ($2'_s$) Mary does not know the (simple) truth $\mathfrak{t}$, to the conclusion that (C'_s) $\mathfrak{t}$ is a non-physical (simple) truth. The reason for this restriction is that, without it, the truth that e is either an electron or not an electron ends up being counted as physical, as would be the truth that nothing is an electron (if it were a truth). The passage from the physicality of a truth to the physicality of its subject-matter would be obscure without this restriction. $\mathfrak{t}$ is a simple truth, since it results from generically quantifying the argument position of the atomic truth that this experience is like this. Henceforth I'll leave this complication tacit.

[17]See, in addition to the discussion above and in the appendix, Lycan 1995 and Stanley and Williamson 2001.

[18]Suppose that the proposition p is of the form $\|o$ is $F\|$. I take it that one finds p to be physical just in case (i) when one entertains the sense of o, one finds its referent to be physical, and (ii) the same for F.

[19]Some property correlated with expressibility might be at work here: I discuss such a property below. Expressibility is somewhat more obvious than this other property, so I suspect that expressibility is doing the real work here.

[20] "We cannot form to ourselves the idea of a pine-apple, without having actually tasted it". See also the citations in Daly 1998.

[21] As I remarked above, I am uncertain whether there are *a priori* connections between obviously physical and reflective/sympathetic concepts.

[22] Jackson (1998) and Chalmers (2002) push lines closely resembling this one.

[23] Example: (pointing at a dog) 'I want a dog that is like that (dog)' to mean 'I want a dog that is similar to that (dog)'/'I want a dog that resembles that (dog)'.

[24] Special thanks to Harold Hodes for incisive comments on the arguments to follow.

[25] Following a suggestion due to Harold Hodes, perhaps resemblance must be conveyed, but what is conveyed is not "first-order" resemblance between my couch and the bearer of the demonstrated instance of redness, but rather "second-order" resemblance between my couches color and the color of that bearer. According to Hodes, this suggestion makes better sense of the fact that utterances of (d), unlike utterances of 'my couch is thus', need not convey that my couch and that bearer exactly resemble in color. But first, I doubt that utterances of 'my couch is thus' require exact resemblance: perhaps the exact-resemblance intuition results from thinking of 'thus' as stressed, which would require exact resemblance. Second, what is demonstrated can be a determinable property, so I also deny that, according to my view, exact resemblance is required. Third, it is unclear what is meant by the suggestion that properties resemble. Bearers of properties resemble one another in virtue of having properties, say scarlet and crimson, which are close to one another in some metric of determinates, say the color solid. So, perhaps if scarlet and crimson resemble one another, they too have properties which are close to one another in some metric of determinates. But what properties could these possibly be? Fourth, it would not be so bad for me were Hodes's suggestion correct, since, if Hodes is correct, what is conveyed must concern a property, rather than an individual. It thus does equally well at replying to Lewis's objection and explaining the specification

phenomenon.

[26]The '%' represents semantic anomalousness.

[27]Thanks to Sally McConnell-Ginet for helpful discussion of this argument.

[28]Thanks to Sydney Shoemaker for pressing me on this example.

[29]Although there are pairs of *wh* forms and proforms which do not match so closely in their pronunciation and spelling, such as 'he'/'she' / 'who' and 'it'/'that'/'this' / 'what'/*'whis', these are plausibly regarded as "irregular" forms, which persist in the face of general rules due to their frequency of occurrence. It should be noted that these forms are also declined, whereas English is generally uninflected.

A defeasible rule can be extracted from this pattern. The phonological realization of the *wh* form and the proform of a particular semantic type differ only in that the former begins with a 'wh' sound whereas the latter begins with a 'th' sound. This rule seems to be at work with predicate proforms and predicate *wh* forms. When 'thus' went out of style to be replaced by 'like that' as the dominant predicate proform, the *wh* form 'like what' became available.

[30]Namely, they can be base-generated in the same positions. In English, a *wh* form is always pronounced at the beginning of the clause, displaced from the positions in which it are base-generated.

[31]Unlike 'thus', 'how' has not gone out of fashion. Thus, English is blessed with an abundance of predicate *wh* forms, and 'I know what it's like' is equivalent to 'I know how it is'.

[32]Something slightly odd happens in generating the predicative reading: the semantic unit 'like what', a predicate *wh*-form, is not being treated as a phonological unit, but has rather become scattered over the sentence. However, this does not indicate that 'like' is meaningless, mere phonological junk, as it is in 'this seminar is on, like, presentism'. If 'like' is absent, the result is 'Mary knows what experiences

of seeing a red thing are', which has a somewhat different meaning: it seems to be appropriately answered with an NP, rather than with an adjective. Compare: 'Mary knows what lions are'/'Mary knows what lions are like'. 'Dauntless hunters' adequately specifies the former; 'frightening' completes the latter.

I will speculate wildly about how this scattering arises. For whatever reason, English has lost its proper predicate proform 'thus'. A language needs a predicate proform, so for whatever reason, the compound 'like this/that' was pressed into this role. Standardly a *wh* form sounds like the proform of its semantic type: see fn. 29. So the predicate *wh* form must become 'like what'. Standard rules of question formation require *wh* to move away from the position it questions. These rules did not bother to check with the rules that formed 'like what' out of 'like this/that', so they do not recognize that 'like what' is a *wh* form; as a result, only 'what' gets moved and 'like' is left behind.

[33]It's also a somewhat opinionated overview, which is made necessary by my operating with a theory of propositions somewhat at odds with that employed in the literature on questions. Aside from these departures, my discussion follows that in Stanley and Williamson 2001.

[34]Contextually-set parameters play an important role at every stage here. In order to keep the presentation reasonably accessible, I'll simply elide this role.

References

Bigelow, John and Robert Pargetter, 1990. 'Acquaintance With Qualia'. *Theoria*, 56:129–47.

Byrne, Alex and David R. Hilbert, editors, 1997. *Readings on Color: The Philosophy of Color*, volume i. Cambridge, MA: The MIT Press.

Chalmers, David J., 2002. 'Consciousness and Its Place in Nature'. In David J. Chalmers, editor, *Philosophy of Mind: Classical and Contemporary Readings*. Oxford: Oxford University Press.

Daly, Chris, 1998. 'Modality and Acquaintance with Properties'. *The Monist*, 81:44–68.

Hardin, Clyde L. 1997. 'Reinverting the Spectrum'. In Alex Byrne and David R. Hilbert, editors, *The Philosophy of Color*, volume i of *Readings on Color*. Cambridge, MA: The MIT Press.

Harman, Gilbert, 1996. 'Explaining Objective Color in Terms of Subjective Reactions'. In Enrique Villanueva, editor, *Perception*, volume 7 of *Philosophical Issues*, 1–18. Atascadero: Ridgeview. Reprinted in Byrne and Hilbert 1997.

Hilbert, David R. and Mark K. Kalderon, 2000. 'Color and the Inverted Spectrum'. In Steven Davis, editor, *Color Perception: Philosophical, Psychological, Artistic, and Computational Perspectives*. Oxford: Oxford University Press.

Hume, David, 1739/1978. *Treatise on Human Nature*. Oxford: Oxford University Press.

Jackson, Frank, 1982. 'Epiphenomenal Qualia'. *Philosophical Quarterly*, 32:127–36.

Jackson, Frank, 1986. 'What Mary Didn't Know'. *Journal of Philosophy*, 83:291–5.

Jackson, Frank, 1998. *From Metaphysics to Ethics*. Oxford: Oxford University Press.

Johnston, Mark, 1992. 'How to Speak of the Colors'. *Philosophical Studies*, 68:221–63. Reprinted in Byrne and Hilbert 1997.

Lewis, David, 1988. 'What Experience Teaches'. *Proceedings of the Russellian Society*, 29–57. Reprinted in Lewis 1999.

Lewis, David, 1995. 'Should a Materialist Believe in Qualia?' *Australasian Journal of Philosophy*, 73:140–4. Reprinted in Lewis 1999.

Lewis, David, 1999. *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press.

Lormand, Eric, in preparation. 'The Explanatory Stopgap'.

Lycan, William G. 1995. 'A Limited Defense of Phenomenal Information'. In Thomas Metzinger, editor, *Conscious Experience*. Paderborn: Schöningh.

Nagel, Thomas, 1974. 'What Is It Like to Be a Bat?' *The Philosophical Review*, 83:435–50.

Stanley, Jason C. and Timothy Williamson, 2001. 'Knowing How'. *Journal of Philosophy*, 98:411–44.

Stoljar, Daniel and Yujin Nagasawa, 2003. 'Introduction'. In Peter Ludlow, Yujin Nagasawa, and Daniel Stoljar, editors, *There's Something About Mary*. Cambridge, MA: The MIT Press.

Williamson, Timothy, 2000. *Knowledge and its Limits*. Oxford: Oxford University Press.