Michael Beaton

What RoboDennett Still Doesn't Know

I: Introduction

Mary, the colour-deprived neuroscientist, embodies perhaps the best known form of the knowledge argument against physicalism (Jackson, 1982; 1986). She is a better-than-world-class[1] neuroscientist. Living in an entirely black-and-white environment, she has learnt all the physical facts[2] about human colour vision. She is supposed to be enough like us to be capable of having the sort of experiences that we would have on exposure to colour, but to be clever enough to know and understand the physical facts about her own colour vision, and to be able to work out all the relevant consequences of the facts which she knows.

The key premise of this form of the knowledge argument is that when Mary is finally released from her black and white captivity and shown coloured objects, she will learn something: namely, what it is actually like to see in colour. Indeed, in Frank Jackson's original paper, he takes it to be 'just obvious' that Mary will 'learn something about the world and our visual experience of it' (Jackson, 1982, p. 130) on her release.

Correspondence:
Mike Beaton, Centre for Research in Cognitive Science, University of Sussex, Falmer, Sussex, BN1 9QH, UK. *Email: M.J.S.Beaton@sussex.ac.uk*

[1] Though perhaps not perfect, of which more later.

[2] I will use phrases such as 'physical facts', 'propositional facts', 'propositional knowledge' etc. more or less interchangeably to refer to the objective knowledge which Mary gains from black and white books, videos and so forth. Jackson (1986) states (or perhaps, claims) that after such an education a clever enough Mary could know 'everything in completed physics, chemistry, and neurophysiology, and all there is to know about the causal and relational facts consequent upon all this' (p.291). In this context, Alter (1998) has talked of the 'discursively learnable' facts and Churchland (1989) talks of those facts which are 'adequately expressible in an English sentence' (p.144). I am happy to accept the standard set-up of the knowledge argument, in which such knowledge exists, is learnable by a clever enough student via the route described, and is, further, contrastable with knowledge such as 'red is like this' which Mary does not gain (at least not directly) from her black and white book learning.

Journal of Consciousness Studies, **12**, No. 12, 2005, pp. 3–25

The following, then, is a simple version of Jackson's original knowledge argument, (all premises refer to Mary's pre-release epistemic status):

(1) Mary knows all the physical facts about colour vision
(2) Mary will learn something about what it is like to see in colour on her release
Presumed corollary:
Mary does not know all the facts about colour vision
(3) Physicalism requires that if Mary knows all the physical facts then she knows all the facts
Conclusion:
Physicalism is false

Premise (2) both implies and is implied by what I will call 'the Mary intuition'. This is the intuition that Mary, in the circumstances described, will still learn something on first seeing a coloured object (equivalently, that there is something that Mary, in the circumstances described, does not yet know, namely what it is like to see in colour).

Jackson himself has presented a clarified form of his argument somewhat along the above lines (Jackson, 1986, p. 293). However Paul Churchland has argued persuasively (Churchland, 1985; 1989; 1998) that every possible form of Jackson's argument requires some equivalent of premise (3) above which only appears to go through because of equivocation on two different senses of the word 'knows'.

The knowledge argument, *qua* argument against physicalism, fails, on Churchland's account, not because Mary learns nothing on her release, but rather because she comes to 'know', in a new way, something which she already 'knew' as a set of propositional facts. The physical nature of this 'new' type of knowledge is something which Churchland addresses in detail, as we will see below.

This is one of two possible physicalist responses to the knowledge argument. On this account, the Mary intuition is fully compatible with physicalism. The other major approach is to argue for the falsity of premise (2): to argue that the Mary intuition is incompatible with physicalism. Such a response amounts to a defence of the validity (though not the soundness) of the knowledge argument: it implies the claim that there is indeed some valid reasoning which shows that Mary's learning something new is incompatible with physicalism, exactly as Jackson originally claimed.

Jackson's own recent rejection of his knowledge argument looks at first like the second kind of response. In his initial retraction (Jackson, 1998), he stated that '*after* the strength of the case for physicalism has been properly absorbed' (p. vii), one is 'reluctantly' (p. vii) led to conclude that 'The redness of *our* reds can be deduced in principle from enough [information] about the physical nature of our world despite the manifest appearance to the contrary that the knowledge argument trades on' (pp. 76–7). More recently Jackson (2003) has stated that 'physicalists are entitled to reject the epistemic intuition' (p. 9) 'that founds the

knowledge argument' (p. 2), namely the intuition 'that it is impossible to deduce what it is like to sense red from the physical account of our world' (p. 17).

It sounds as if Jackson is preserving the validity of his knowledge argument by rejecting its second premise, but this is not what is going on. In fact, Jackson still accepts the truth of what I have called the Mary intuition, even while denying what he calls the epistemic intuition. Jackson still believes that Mary 'would learn what it is like to see red' (p. 3) on her release (indeed he continues to treat this as an obvious fact, in need of no defence). What he now denies is 'that this would be learning something about the nature of the world' (p. 3). On Jackson's revised account, Mary will learn no new fact about the world, but will instead gain a new kind of representation; one with the right properties to account for the 'immediacy, inextricability, and richness' of seeing red, and one which additionally grants her the ability to 'recognise, imagine and remember' red (Jackson, 2003, p. 26).[3]

For Jackson, the truth of the Mary intuition remains obvious. The mistake in the knowledge argument was to credit the epistemic intuition: to conclude from Mary's coming to know what it is like to see red after her release, that she thereby comes to know any new fact about the world. Jackson's revised position effectively leaves the knowledge argument exactly where Churchland left it, with true premises, but nevertheless invalid due to equivocation on two senses of 'knows'.

If we accept these arguments, can we consider interesting discussion on the knowledge argument closed? Apparently not, for the above, seemingly straightforward, physicalist consensus on the logical status of the knowledge argument remains radically different from the position held by Daniel Dennett (who is, of course, another die-hard physicalist).

Dennett's position is made clear in his new paper on the subject, 'What RoboMary Knows' (Dennett, 2005).[4] For Dennett, 'most people's unexamined assumptions imply dualism' (p. 107; for which, in context, read 'the Mary intuition is incompatible with physicalism'). The explicit objective of Dennett's new paper is to show that the Mary intuition is an anti-physicalist confusion. He aims to demonstrate — for the benefit of those philosophers who doubt that it can be done — how Mary '*figures out* exactly what it is like to see red (and green, and blue)' (p. 122).

But why should Dennett believe that most people's unexamined assumptions imply dualism? Or that philosophers need to understand why the Mary intuition is false in order to understand how physicalism can be true? He must still believe that there is some logically valid form of the knowledge argument, implying a fundamental incompatibility between the Mary intuition and physicalism, despite all the arguments to the contrary. There is, quite simply, no reason to

[3] As Jackson himself points out, he has thus come to adopt the rejection of his knowledge argument originally employed by Nemirow (1980) and Lewis (1983).

[4] The present paper was originally written in response to an article of Dennett's which is to appear in a collection of new papers on phenomenal knowledge (Alter and Walter, 2006), and which is currently available online at *http://ase.tufts.edu/cogstud/papers/RoboMaryfinal.htm*. Dennett's paper is now in print in slightly modified form as Chapter 5 of his 'Sweet Dreams' (Dennett, 2005). All subsequent quotes from Dennett refer to Ch.5 of *Sweet Dreams* unless otherwise indicated.

deny premise (2) of the knowledge argument, unless you believe that the reasoning in the knowledge argument, or something very like it, is valid. If you accept that premise (1) makes a workable thought experiment, if you want to preserve physicalism, and if and only if you also think that physicalism has some entailment very like that claimed for it in premise (3), then (and only then) will you need to deny premise (2).

This is precisely Dennett's position. So has he simply missed the equivocation on 'knows' from which, Churchland has claimed, all forms of premise (3) suffer? The actual situation is more complex than that, and more interesting. The explicit aim of Dennett's new paper is to show that Mary will necessarily be able to come to know what it is like to see in colour, if she fully understands all the physical facts about colour vision. I believe we can establish that Dennett's line of reasoning is flawed, but the flaw is not as simple as an equivocation on 'knows'. Rather, it goes to the heart of functionalism and hinges on whether or not Dennett is correct to claim that there is 'no fact of the matter' (Dennett, 1988; 1991; 1994; etc.) about what subjective experience consists in.

II: The Blue Banana Alternative

Dennett's previous major position statement on the knowledge argument occurred in his book *Consciousness Explained* (Dennett, 1991, pp. 398–401). There, he first outlined in print what he believes to be a perfectly legitimate alternative ending to the Mary story. Instead of experiencing 'surprise and delight' (Graham and Horgan, 2000, p. 72) on being released from her room and first seeing coloured objects, something quite different happens. Mary's captors decide to trick her, and the first coloured object they allow her to see is a blue banana. Dennett doesn't explicitly state as much, but presumably Mary's captors are expecting Mary to say to herself something like, 'Ah, so that is what yellow looks like!' However Mary isn't fooled for a moment, she takes one look at the blue banana and says, 'Hey! You tried to trick me! Bananas are yellow, but this one is blue!' and further 'I was not in the slightest surprised by my experience of blue (what surprised me was that you would try such a second-rate trick on me)' (Dennett, 1991, pp. 399–400).

Dennett states that students and professional philosophers alike have had considerable problems with his alternative ending to the story (Dennett, 2005, p. 106). What is he saying? Is he seriously trying to claim that Mary has 'figured out' what it is like to see in colour without ever having seen anything coloured? That is, of course, exactly what he is trying to claim. And he is not just stating that Mary will know enough about her own physical reactions to colour to be able to recognize them when they first occur, and so work out what colour she has seen. He is, rather, taking the following much stronger position: that knowing as much about your own reactions in advance of the fact as Mary does is logically equivalent to knowing what it is like to see colour in advance of the fact.

He explicitly states that he knows of no 'distinction … between knowing '*what one would say and how one would react*' and knowing 'what it is like'. If

there is such a distinction, it has not yet been articulated and defended, by [anyone], so far as I know' (Dennett, 2005, footnote 3).

To many, of course (even to those who hold to the truth of some form of physicalism) this current, clear and explicit statement of position by Dennett will itself seem extreme. This is why he has felt compelled to return to the fray, and to attempt to 'convince a few philosophers' (Dennett, 2006) that his position might be correct after all.

III: Introducing RoboMary

Dennett's chosen weapon for his final attack on the knowledge argument is RoboMary, a perfected robot neuroscientist. Dennett uses RoboMary because he needs to discuss the physical details of her behaviour and thought processes at a level of detail not currently available to human neuroscience. Using RoboMary he hopes to show, by analogy, how a human-like Mary could also come to know what it is like in advance of the experience.

I am happy with this approach, and agree with Dennett that a physicalist account of what is really going on in the Mary thought experiment will require a discussion of the physical details of the 'agent' under discussion. As Dennett says:

> If materialism is true, it should be possible ('in principle!') to build a material thing — call it a robot brain — that does what a brain does, and hence instantiates the same theory of experience that we do (Dennett, 2006).

And further:

> Those who rule out my scenario as irrelevant from the outset are not arguing for the falsity of materialism; they are assuming it (p. 125).

Dennett wants to make sure that RoboMary is a well constructed and well labelled 'intuition pump'. He succeeds admirably. In fact, once I have summarized here Dennett's key 'knobs' and 'settings' for RoboMary, she will make an ideal subject on which to attempt some 'cooperative reverse-engineering' (p. 122) of my own.

There are two major models of RoboMary, either of which, it is argued, can come to know what it is like to see in colour in advance of the experience. As Dennett outlines these two versions of RoboMary he considers and refutes many possible objections to his account. On many, indeed most, of these points I am fully in agreement with Dennett. Therefore I will only give an outline of the key facts about RoboMary, omitting the several objections to his story that Dennett successfully addresses.

IV: Unlocked RoboMary

The basic RoboMary model is (for reasons presumably lost in the mists of sci-fi time) a standard Mark 19 robot. The easiest thing to do will be to quote directly the key points from Dennett's story about her:

> 1. RoboMary is a standard Mark 19 robot, except that she was brought on line without color vision; her video cameras are black and white, but everything else in her hardware is equipped for color vision, which is standard in the Mark 19.
>
> 2. While waiting for a pair of color cameras to replace her black-and-white cameras, RoboMary learns everything she can about the color vision of Mark 19s. She even brings colored objects into her prison cell along with normally color-sighted Mark 19s and compares their responses — internal and external — to hers.
>
> 3. She learns all about the million-shade color-coding system that is shared by all Mark 19s.
>
> 4. Using her vast knowledge, she writes some code that enables her to colorize the input from her black and white cameras (à la Ted Turner's cable network) according to voluminous data she gathers about what colors things in the world are, and how Mark 19s normally encode these. So now when she looks with her black-and-white cameras at a ripe banana, she 'sees it as yellow' since her colorizing prosthesis has swiftly looked up the standard ripe-banana color-number-profile and digitally inserted it in each frame in all the right pixels.
>
> 5. She wonders if the ersatz coloring scheme she's installed in herself is high fidelity. So during her research and development phase, she checks the numbers in her registers (the registers that transiently store the information about the colors of the things in front of her cameras) with the numbers in the same registers of other Mark 19s looking at the same objects with their color camera eyes, and makes adjustments when necessary, gradually building up a good version of normal Mark 19 color vision.
>
> 6. The big day arrives. When she finally gets her color cameras installed, and disables her colorizing software, and opens her eyes, she notices … nothing. In fact, she has to check to make sure she has the color cameras installed. She has learned nothing. She already knew exactly what it would be like for her to see colors just the way other Mark 19s do (pp. 122–5).

For what it is worth, I buy into this story. There don't seem to me to be any interesting reasons why RoboMary can't do what Dennett claims, above, that she can do. And if she can indeed do the above then she would indeed come to know what it is like to see in colour in advance of the experience. But an objection that Dennett considers concerning his step 4 is the crucial one, in terms of relating the story of unlocked RoboMary to the story of Mary. The question is, is unlocked RoboMary cheating or not when she writes directly to her colour coding registers? Perhaps, as Dennett himself says, RoboMary's colourising system is simply the 'robot version … of transcranial magnetic stimulation' (p. 124): cheating in the sense of using a non-surprising way of coming to know what it is like, which doesn't truly involve deducing what it is like from the facts one knows. Or perhaps we should accept that 'RoboMary is entitled to use her imagination, and that is just what she is doing — after all, no hardware additions are involved' (p. 124).

Dennett is happy to vary this setting in both directions. For reasons related to the above point about imagination, my understanding is that Dennett thinks there is no truly principled reason to rule out even this unlocked version of RoboMary

as a counter-example to the Mary intuition. (I will argue below that there is, in fact, a principled reason to rule unlocked RoboMary's route to coming to know what it is like as cheating.) Nevertheless Dennett is happy to take on board this objection, and to consider next a much more challenging version of the RoboMary story.

V: Locked RoboMary

Following Dennett, 'let's turn the knob and consider the way RoboMary must proceed if she is prohibited from tampering with her color-experience registers' (p. 126). The use of a robot instead of a human in the thought experiment once again pays dividends. As Dennett says, we have no idea how 'Mary could be crisply rendered incapable of using her knowledge to put her own brain into the relevant imaginative and experiential states' (p.126), but we can easily describe something equivalent for RoboMary. We can put a software system in place which automatically converts all the colour values in RoboMary's visual array to black and white (or rather, greyscale) values before any further processing takes place. Now let's put unbreakable software security on this system. Suddenly RoboMary really can't 'imagine' herself into any normal colour vision state. She can't even create colour 'phosphenes' (one objection to the original Mary story) by any robot equivalent of rubbing her eyes. The only way her colour registers can ever come to contain any usable colour values is for the software security system to be disabled which, let us assume, requires a hardware change and so can be treated as unambiguous cheating.

Surely then there is no way for RoboMary to deduce what it is like to see in colour, is there? Oh yes there is, says Dennett:

> This doesn't faze her for a minute, however. Using a few terabytes of spare (undedicated) RAM, she builds a model of herself and *from the outside, just as she would if she were building a model of some other being's color vision*, she figures out just how she would react in every possible color situation (p. 126).

This is supposed to be pure heterophenomenology. For Dennett, there can be no distinction between the full facts about 'what one would say and how one would react' and the full facts about 'what it is like'. Thus, if RoboMary can indeed build such a model, she can indeed come to know what it is like. QED.

But the preceding is a reconstructed abbreviation of Dennett's argument. Let's follow the actual details of the story which Dennett gives. Rather than mix and match direct and indirect quotation, I will paraphrase this section of Dennett's argument (pp. 127–8). Imagine, says Dennett, a situation in which (locked) RoboMary is shown a ripe tomato. She can see it and touch it and find out all about its bulginess and softness. She can also consult an encyclopaedia to find out exactly what shade of red it would be, if only her colour registers were unlocked. RoboMary will react in various ways to this stimulus, resulting in some complex, internal, grey tomato experiencing state, state A. But at the same time, she can feed into her internal model of herself the true red colour values which she knows she would have seen if her colour vision equipment was normal

for Mark 19s. So her model will go into a different complex state, a red-tomato-experiencing state, state B. This should be fine: the model RoboMary doesn't have to be 'locked', just because RoboMary is. She knows all about how she would work if she was not locked, and so she should be able to build and operate an unlocked model just as Dennett describes. So now, returning to direct quotation, locked RoboMary compares state A with state B and:

> being such a clever, indefatigable and nearly omniscient being – makes all the necessary adjustments and *puts herself into state B* (p. 128).

Dennett is at pains to point out that state B really isn't an illicit state in the sense in which direct tampering with colour registers is an illicit state. State B is the state that Mary would have gone into if she had had the colour experience, even though she hasn't in fact had it: she isn't making herself experience colour (cheating) she is making herself be as she would be if she had experienced colour (not cheating).[5]

I am prepared to buy into this story, too. I accept that locked RoboMary could find such a state and put herself into it. But I don't accept that RoboMary has told us about what must be true of an agent who knows what it is like in the way that we do. To explain why, the first thing we must do is try to be as clear as possible about what we mean by knowing what it is like within human cognitive architecture.

VI: The Churchland-Lewis Account

To obtain the details of human cognitive architecture on which I wish to draw, I will briefly recall two very well known accounts of how it might be that a consistently defined and completely physical Mary could come to know all the facts about colour vision and still not know what it is like to see in colour.

Paul Churchland and David Lewis were two of the first authors to present an 'ability' or 'knowledge how vs. knowledge that' response to Frank Jackson's knowledge argument.

Lewis' distinction between these two forms of knowledge occurs in a postscript (Lewis, 1983) to an earlier paper (Lewis, 1980). In his postscript Lewis states that 'The most formidable challenge to any sort of materialism and functionalism comes from the friend of phenomenal qualia.' Lewis details the nature of this perceived challenge by presenting his own version of the knowledge argument which parallels Jackson's, using the taste of Vegemite instead of the visual experience of colour. He concludes:

> We dare not grant that there is a sort of information we overlook; or, in other words, that there are possibilities exactly alike in the respects we know of, yet different in some other way. That would be defeat. Neither can we credibly claim that lessons in

[5] Dennett draws an instructive analogy here with Swamp Mary (another character whom Dennett introduces, whilst suppressing his 'gag reflex' and 'giggle reflex'; p.120). You may be happy to infer the details for yourself, or you may wish to refer to Dennett's paper, but I think that his point goes through.

> physics, physiology, … could teach the inexperienced what it is like to taste Vegemite.

That is to say, of course, that (a) epiphenomenal, or otherwise non-physical, qualia must be rejected, but nevertheless that (b) as far as Lewis is concerned the Mary intuition (in this case, the Vegemite intuition) is correct: someone who has not tasted Vegemite cannot know what it is like, however much they know of the physical facts.

Lewis concludes that the proper resolution must lie in the realization that: 'knowing what it's like is not the possession of information at all', rather it is the 'possession of abilities … to recognize, … to imagine, … to predict one's behaviour by means of imaginative experiments'.

He goes on to flesh out the kind of thing he is thinking about:

> Imagine a smart data bank. It can be told things, it can store the information it is given, it can reason with it, it can answer questions on the basis of its stored information. Now imagine a pattern-recognizing device that works as follows. When exposed to a pattern it makes a sort of template, which it then applies to patterns presented to it in future. Now imagine one device with both faculties… There is no reason to think that any such device must have a third faculty: a faculty of making templates for patterns it has never been exposed to… If it has a full description about a pattern but no template for it, it lacks an ability but it doesn't lack information. (Rather, it lacks information in usable form.) When it is shown the pattern it makes a template and gains abilities, but it gains no information.

'We might', Lewis suggests, 'be rather like that.'

Indeed we might.

The details of Churchland's account occur in his second response to the knowledge argument (Churchland, 1989, pp.145–7). Though considerably more detailed than Lewis', Churchland's is an account of essentially the same distinction. As Churchland says, 'modern cognitive neurobiology already provides us with a plausible account of what the difference is' between 'knowledge by description' and 'knowledge by acquaintance'. He points out that in all trichromatic creatures 'color information is coded as a pattern of spiking frequencies [within] the optic nerve'. This 'massive cable of axons' projects to the 'lateral geniculate nucleus (LGN)' and thence 'to V1, V2, and ultimately to V4, which area appears to be especially devoted to the processing and representation of color.' The model of visual information processing that Churchland then appeals to is one which assumes that the 'representation of familiar colors … consist[s] in a specific configuration of weighted synaptic connections meeting the millions of neurons that make up area V4.' This 'configuration of synaptic weights partitions the [abstract] activation-space of the neurons in area V4 … into a structured set of subspaces, one for each prototypical color.' New patterns of input from the eye can then be categorized accordingly. 'In such a pigeonholing, it … appears, does visual recognition of a color consist.'

Churchland concludes:

> This distributed representation is not remotely propositional or discursive, but it is entirely real. All trichromatic animals have one, even those without any linguistic

> capacity. It apparently makes possible the many abilities we expect from color-competent creatures: discriminations, recognition, imagination, and so on. Such a representation is presumably what a person with Mary's upbringing would lack, or possess only in … incomplete form. There is thus more than just a clutch of abilities missing in Mary: there is a complex representation, a processing framework that deserves to be called cognitive… There is indeed something she 'does not know.' Jackson's premise … is thus true on … wholly materialist assumptions.

In order to provide the details in the above account Churchland admits that he has 'momentarily' put 'caution and qualification … aside'. In other words, Churchland believed, in 1989, that the state of neuroscientific knowledge was such as to allow us to know perfectly well that there was such a story to be told about the human brain, but to have to guess at many of the details. The precise details of Churchland's account do not, I hope, matter because this overview of the situation remains accurate: we still know that there is such a story to be told and we still have to guess at many of the relevant details.

VII: Knowing What It Is Like

Churchland and Lewis seem to be pretty much in agreement about what happens in Mary's brain on her release: some lower level colour processing circuitry comes to be configured due to exposure to colour. From then on, this circuitry enables Mary to recognise colours to which she has been exposed, and to remember and imagine colour. But this agreement on what happens to Mary does not lead to a common account of the state of knowing what it is like.

Paul Churchland suggests (Churchland, 1985) that after exposure to colour, Mary knows directly (non-inferentially) that her visual system is in a certain physical state. He presents this as a direct parallel to his account of perception more generally, within which perceiving, say, the temperature of an object means knowing non-inferentially that the object has a certain distribution of energy across micro-states. In either case, the subject doesn't (without additional tutoring and practice) know the relevant facts under the relevant scientific concepts, nevertheless those are the facts which the subject actually (opaquely) knows.[6] On this account, it is only after exposure to colour that Mary can directly introspect some particular physical fact about her brain. Before exposure to colour, she knew (in terms of information expressed as propositions) what state her brain would go into. After exposure to colour she additionally knows *the same fact* directly, 'by acquaintance'.

On David Lewis' account, on the other hand, Mary does not gain any new knowledge at all: not knowledge of the world, not knowledge by acquaintance of a particular brain state. She simply gains certain abilities: the abilities to 'recognise, imagine and remember' which Jackson himself now takes to be constitutive of knowing what it is like. Nemirow (1990) explains the work that the phrase 'knowing what it is like' is doing on this account: he suggests that the

[6] Churchland further suggests that — for both perception and introspection — it is possible to learn to perceive these facts directly (noninferentially) under the relevant scientific concepts.

sub-expression 'what it is like' of the phrase 'knowing what it is like' is 'a 'pseudosingular term' — an expression that has the grammatical form of a singular term, but, on analysis, does not even purport to refer' (p. 494). Instead, the entire phrase 'x knows what it is like' is a locution which means that x has the relevant abilities.

I wish to propose yet another analysis of the state of knowing what it is like, based on the same story about the physical changes in Mary. I will disagree with Churchland's analysis. I don't think that Mary learns anything new about her brain, rather, she gains a new way of thinking about the world. I will essentially agree with Nemirow, Lewis and Jackson, though I'll try to be somewhat more explicit about how it is that Mary's new subpersonal recognitional abilities grant her the full blown personal level abilities which she gains.

My central claim will be the following:

The state of knowing what it is like to perceive X is the state in which one's more abstract capacity to reason about X is supported by lower level sensory processing apparatus which discriminates and responds to X. (K)

On this account, prior to her release Mary knew what red was, because she could tell you all about human colour vision, and could predict exactly which things, under what conditions, humans would classify as red. Moreover, she could tell you all about the innate predispositions to treat red in a certain way (not 'conceptually', but purely behaviourally) which she no doubt shares with her primate ancestors (Humphrey and Keeble, 1978).

The concept of red which Mary has, pre-release, is what I have called propositional knowledge. She knows this information about the physical universe in a form which relies on no particular type of grounding of her knowledge. She has learnt about colour vision from black and white books and videos (and perhaps experiments carried out by her, as long as she only ever sees what she is doing, and what her actions result in, in black and white). Her 'objective', propositional knowledge of red does have to be grounded in ostensive definition somewhere. What makes her knowledge objective is that a different agent, with different innate abilities, could have learnt the equivalent facts about light, and about human colour vision. But the knowledge argument requires that Mary start off with no opportunity to use sensory apparatus which can pick out red stimuli directly — more precisely, it requires that she does not have the correctly configured sensory apparatus that humans normally have, that picks out red for us, and which supports and enables higher level, more abstract processing about red as experienced.

We would be wrong to think that, when Mary is released, just gaining the correctly configured sensory apparatus is sufficient for her to know what it is like to see red. Just picking out red — in V4 say — without any connection to the brain regions responsible for additional, more conceptual abilities, would be the equivalent of blindsight. In order for Mary to come to consciously see red, she has to be able to *report* that she is seeing red now, to be able to choose to

remember and imagine it at will, and to act on her imagination as appropriate, after having seen it.[7] It seems highly likely that these more complex, conceptual abilities are functions of the associative and frontal regions of the human brain (Fuster, 2004), regions which are functionally distinct from lower level sensory cortex in the important sense that sensory cortex can continue to effectively carry out the vast majority of its tasks without the presence of higher brain regions, whilst the converse is not true (Laureys, Faymonville *et al*., 2004).

On the present view then, on exposure to colour, Mary gains a new configuration in her sensory cortex (specifically in the region dedicated to the processing of colour), but she additionally gains a new neural configuration in her associative and/or frontal cortical regions. This additional configuration corresponds to Mary's having gained a new concept, a concept which I will gloss as '*red_as_experienced*'. Mary can still think in terms of the propositional, objective, concept of red which she previously possessed; a concept which must have been grounded somehow. But she now possesses a new concept, of red as experienced, grounded in the very sensory apparatus which enables her to detect and respond to red stimuli.

What we now need to consider, following Dennett, is whether or not an agent who knows as much as Mary would be able to use her highly detailed and advanced propositional conception of 'red' to derive the specific grounded concept 'red_as_experienced'.[8]

VIII: What Physicalism Requires

For convenience, let's recap, with a quick and simple version of the knowledge argument:

(1) Mary knows all the physical facts
(2) Mary does not know what it is like
(3) Physicalism says that if you know all the physical facts then you know everything
Conclusion:
Physicalism is false

How should a physicalist respond?

Most physicalists, including Jackson (now), Nemirow, Lewis and Churchland have been prepared to accept that there is some distinction between the type of knowledge which Mary has, pre-release, and the type of knowledge which she gains on her release. Some physicalists have argued that Mary gains a new ability but does not thereby come to know any fact — not even an old fact in a new way; other physicalists have argued that Mary gains a new type of knowledge of an old

[7] Such abilities do not, I think, require language (cf. Cowey and Stoerig, 1995).

[8] We may note, in passing, that if Mary always can do this, then a Mary without any colour vision system at all would also be able to do the same thing. It doesn't take many additional steps to claim that such a Mary would indeed be able to work out exactly what it was like to be a bat, for instance. At issue is the question of whether or not physicalism requires this.

fact. The important point here is that both these responses accept that it is possible for Mary to know all the physical facts and, at one and the same time, not to know what it is like.

Surprisingly, perhaps, even Dennett accepts this.

In either version of Dennett's story, RoboMary has to *do something* in order to come to know what it is like. She either has to adjust her colour registers, or she has to work out some special state, *state B*, and put herself into it. She's never just *automatically* in state B, as soon as she's finished learning all the facts. So pre-release RoboMary is like this: if you ask her what it is like to see ultramarine, say, she says 'I don't know, but I can work it out. Hold on a minute [or a second, or a picosecond] … Ok, there we are! Now I know.'

Dennett is happy to accept that RoboMary knows all the physical facts. I believe he is also happy to accept that what it is like to see red is not something which Mary *automatically* knows, just in virtue of premise 1. But he thinks that physicalism requires that Mary be able to *work out* what it is like to see red; that believing otherwise is an anti-physicalist confusion. Why? The version of the knowledge argument which Dennett must be using, the only version whose rejection requires Dennett's arguments, is the following:

(1) Mary knows all the physical facts
(2) Mary cannot work out what it is like
(3) Physicalism requires that if you know all the physical facts, you can work out what it is like

Conclusion:
Physicalism is false.

If you wish to preserve physicalism under this argument, and you accept premises (1) and (3), then you have to reject premise (2). Conversely, if you accept premise (1), and you wish to preserve physicalism, you still have no reason whatsoever to reject premise (2) unless you think that premise (3) is true.

Dennett, of course, does think that premise (3) is true. But why? To be clear about the logical status of premise (3), we have to think about what *might* and what *must* be true of agents who know as much as Mary.

I am not particularly interested in what might be true of agents in 'ectoplasmic' worlds. Let's talk only about universes such as ours (I hope) in which all true facts supervene[9] on the *physical* state of the universe (the state as it would be described, if we had that much knowledge, in terms of the completed laws of physics). We can then ask, what might and what must be true of agents who know as much as Mary, in such purely physical universes? I will say that some predicate is *necessarily* true of such an agent if it must be true of every agent which could possibly be built, consistent with the laws of physics, who knows as much as Mary. I will say that some predicate is *possibly* true of such an agent, if that predicate can be true of an agent who knows that much — consistent with the laws of physics — but doesn't have to be.

[9] This supervenience relationship means, quite simply, that you can't change any fact (of any type) without changing some physical fact. If it's true, it gives (most) physicalists what they want.

Thus, I would claim, it is *necessarily* true that Mary can work out what 2 + 2 comes to, but it is only *possibly* true that (for instance) Mary's brain has built in to it a transcranial magnetic stimulation machine, which she can operate at will, which results in coloured visual phosphenes.

Now, for Dennett's arguments to work, it needs to be the case that Mary can *necessarily* work out what it is like to see red.[10] If she can only *possibly* work this out (if some agents who know that much can work it out, but some other agents who know that much cannot), then Dennett's argument is flawed. At best, RoboMary might lead one to accept that belief in the Mary intuition is belief that Mary has one physically possible type of architecture rather than another, which is not an anti-physicalist position at all. At worst (for Dennett's current position) there may be a good reason to believe that we were supposed to be thinking about an agent with the 'can't-work-it-out' architecture all along. If this were so, the Mary intuition would be better than equally as physical as its denial: it would be the correct intuition to have had about Mary all along.

IX: RoboDennett

I have argued that the key question, which determines whether or not the Mary intuition is compatible with physicalism, is whether or not an agent who knows as much as Mary can necessarily use that knowledge in order to come to know what it is like to see red, if she so chooses.

Now, I will argue that there is nothing in the set up of the knowledge argument which *requires* that Mary be able to do what Dennett's RoboMary does. On the contrary, it will be possible to describe a perfectly physically well defined robot agent who can know quite as much as Mary, or RoboMary, but who remains genuinely unable to come to know what it is like, despite mastering all the abilities that Mary is granted by the first two premises of the knowledge argument.

In order to regiment the discussion we need, finally, to be clear about what we mean by cheating in the context of the knowledge argument. The correct way to proceed is as follows:

> When considering an agent trying to achieve what RoboMary achieves, in the context of the knowledge argument, the agent should be considered to be cheating if it uses abilities other than those entailed by the hypotheses of the knowledge argument.

I have already suggested, in the introduction to this paper, what these abilities are. The agent in question *must* be quite like us, for she must be capable of knowing what it is like to see red in the same way in which we do. Premise (2) requires this — we all grant that, after normal exposure to red, Mary will know what it is like to see red in the same way we all do.

On the analysis proposed above, this means that Mary must have some low level colour processing circuitry which can pigeonhole and then re-identify

[10] We are talking about an A-grade student here, one who will not miss, or misunderstand, consequences of what she knows. As such, and as I will emphasise below, Mary necessarily can get very close.

coloured stimuli when exposed to them. It also means that the circuitry (be it neural or electronic) which enables Mary's more abstract reasoning must be connected to this colour processing circuitry in such a way that when Mary has seen colour, this fact will be integrated with the rest of her more abstract reasoning, granting her the concept '*red_as_experienced*'.[11]

Premise (1) requires that Mary's abstract reasoning powers be much better than ours. She knows everything there is to know about how her own colour vision works. Moreover, she can work out any relevant consequences of what she knows. We should be wary of granting Mary perfect reasoning powers, but I don't believe that we need to. What we need to allow is that anyone trying to show just what Mary can do, can help themselves to any particular reasoning process, by Mary, based on her vast knowledge — but *only* in terms of reasoning from propositionally expressed knowledge to more propositionally expressed knowledge.

This, of course, is the key move, but it does not, yet, establish the falsity of Dennett's position, for, as we will see, there are very good reasons (quite the best reasons, in fact) for thinking that these abilities alone *are* sufficient for creating a *bona fide* state of knowing what it is like.

Using the above limitations, I will define a new robot. I will, of course, name him RoboDennett. RoboDennett is extremely intelligent, and he knows an awful lot — quite as much as Mary, or RoboMary, in fact. The only difference between RoboDennett and RoboMary (if indeed there is a difference) is that RoboDennett has no abilities which are not granted to him by the premises of the knowledge argument.

RoboDennett is, of course, the agent whom we should have been imagining all along, in the context of the knowledge argument. If the Mary intuition is true, of him, then the Mary intuition is not just compatible with physicalism, it is the *correct* intuition to have about someone who starts off like one of us, and who is only changed as little as possible in order to come to know as much as Mary knows. This remains so *even if* there are other physically possible agents, such as RoboMary, who can use all their knowledge to come to know what it is like prior to exposure to colour.

RoboDennett, of course, is very like RoboMary. RoboMary certainly *has* the abilities which I have granted to RoboDennett. The only substantive question is whether or not she exceeds them.

X: RoboDennett and Unlocked RoboMary

I said before that there were principled reasons for declaring that unlocked RoboMary was cheating. You will recall that she works our what colour values should be in her low level colour circuitry, and then simply puts them there. Of

[11] The remainder of this paper is written in terms of a particular functional analysis (K) of knowing what it is like. I believe Dennett would be wrong about RoboMary for the reasons expressed in Sections VIII–XI on any functional account, but I do not have an argument to establish that the conclusions of Section XII follow if K and similar accounts are rejected.

course she can work out what the colour values should be, but there is no reason to think that we humans have the ability to configure our low level colour processing circuitry in the way unlocked RoboMary does, just by thinking about it, in advance of any exposure to colour. And I believe that there is no argument which says that an agent who knows as much as Mary somehow automatically gains the ability to do this. Apologies for having only shifted the burden of proof, but I think I have shifted it quite far. Lacking an argument for the *necessary* presence of this additional ability, unlocked RoboMary really was going beyond her legitimate powers of imagination, she was doing something which we cannot do with our imaginations, and something which increasing our reasoning powers up to the level of Mary's would not enable us to do. She was cheating.

XI: RoboDennett and Locked RoboMary

Now, I don't think Dennett has missed the central point I am making. He is less explicit about it than I have tried to be, but he recognises that what he actually needs to show is that any agent who has mastered all Mary's knowledge must necessarily be able to use that knowledge to come to know what it is like. Further, I think this is precisely what he believes he has shown, using locked RoboMary. As we look in detail at Dennett's reasons for believing that the Mary intuition is fundamentally unphysical, we will see that what locked RoboMary does is indeed, by Dennett's lights, a completely general route to coming to know what it is like, a route which would be available to any agent who knows as much as Mary and can work out the consequences of what she knows.

For most of the steps on locked RoboMary's path to enlightenment, I am in full agreement with Dennett. Nevertheless, I believe that RoboMary does not correctly represent the entailments of physicalism. The final step (and only the final step) which locked RoboMary takes is a perfectly physical move, but it is a step which Dennett should not have allowed her, for it is a step which is not available to RoboDennett.

It is no accident that locked RoboMary's route to coming to know what it is like involves working out exactly what she would say and how she would react on exposure to colour. What she has actually done, just by thinking hard, is to create a simulation of herself.

This is step which RoboDennett *can* take, even without the explicit provision of spare, undedicated RAM and processing power.

Imagine that you, yourself, knew everything about how a pocket calculator works (not the atoms or the quarks, just the registers, the CPU instruction set, and the relevant connections to the keys and the LCD display). Is it plausible that, once you knew all this, you could do without a pocket calculator? Of course not, for you are too human. You would make mistakes sometimes, as you tried to work out what the calculator would do, and even if you were very careful, and did get the answers right, you would be much slower than the calculator.

But to think that RoboDennett would still need a calculator, once he had put his mind to understanding one, is indeed to make precisely the mistake which

Dennett accuses us all of making with regard to Mary. For RoboDennett is much better than us (as, indeed, is RoboMary, and presumably Mary too). Once he has put his mind to understanding a pocket calculator, it will be immediately obvious to him what the result would be of calculating sin(37 /5)^6 (for instance).[12] That is to say, these agents are good. Very good. And, crucially, they are all supposed to be equally good even at the vastly more complex task of understanding themselves.

Are we still within the bounds of sense here? Is it possible to make any meaningful statements about an agent who is supposed to be (a) in some relevant way, human-like, but (b) to know as much, and be as good at using that knowledge, as Mary, RoboMary or RoboDennett are supposed to be? Yes, I believe so, though we have to steer carefully in these waters.

In the example of the calculator, above, RoboDennett's understanding of the calculator becomes good enough for him to do away with the actual calculator if two crucial conditions obtain:

(i) His understanding is so good that it is functionally isomorphic to (the relevant level of organization of) the calculator itself.
(ii) He can operate this functionally isomorphic understanding at least as fast as the calculator itself.[13]

A recent paper by Adams and Aizawa (2001) offers the opinion 'Philosophers these days seem not to appreciate that isomorphism is a relatively weak relation'. I wish to claim that, on the contrary, isomorphism is an exceedingly strong relation. Something physical which is fully, counterfactually (Chalmers, 1994; Chrisley, 1994), functionally isomorphic to a particular definition of a calculator is, in a good sense (quite the best sense, in fact) a calculator. I take it that I am with Dennett on this.

And I accept that RoboDennett can indeed perform such a simulation of himself. As such (and again, I take it that I am with Dennett on this) what RoboDennett can do is generate a *bona fide* state of knowing what it is like. On this very strong functionalist account, RoboDennett has actually created an agent which knows what it is like. It is living in a virtual world, but it wouldn't necessarily know that this is the case (Chalmers, 2003); it is up to the real RoboDennett to decide whether or not to make this information available to the simulation.

At this stage, though, the state of knowing what it is like is a state of the simulation, not a state of the simulating agent. Even on Dennett's account, to come to know what it is like, locked RoboMary has to do something above and beyond creating this simulation. She has to work out the relevant aspect of the state of the

[12] It's approximately 0.74, and I don't happen to know how many decimal places were on the calculator which RoboDennett was thinking about.

[13] Speed of simulation *is* important, here. We will look later at what heterophenomenology requires. If it turns out that there's any fundamental reason why RoboMary's simulation of herself is necessarily slower than the real thing, then we've got a behavioural distinction right there between a RoboMary who really knows what it is like and RoboMary who is just working out how to behave as if she knew what it is like, using a simulation.

simulation (Dennett's state B), and then she has to *put herself into that state*. It is this step which RoboDennett cannot take. He can simulate himself as well as he likes,[14] but that's it.

If the state of knowing what it is like to see in colour is indeed the state in which a low level colour processing system is playing the right causal role of enabling certain grounded conceptual abilities, then we need to ask whether RoboDennett can make his low level colour processing system play that causal role. If he cannot, we need to ask whether he can make *anything else* play the relevant causal role. If he can do neither of these things, then he simply will not be in the state of knowing what it is like, despite all his knowledge.

The first option above is unlocked RoboMary's route to coming to know what it is like. We have already rejected it as cheating, in quite a precise sense, and we need not consider it again. RoboDennett cannot do it.

What about trying the second option, of getting something else to play the relevant causal role? Again, RoboDennett can come tantalisingly close. He can't tamper with his actual colour categorisation system, but he can think very hard, and thereby bring into existence a perfectly good simulated colour categorisation system (indeed, one which is as it would be if he had seen colours). Now all he has to do is to put *that* simulation into the right causal relationship with those parts of his brain which enable his propositional reasoning abilities. Again, RoboDennett can do everything except the last step.

The ability to think very hard requires that an agent have very advanced, reason respecting transitions between its many and various thoughts. As we've mentioned, it also requires that there be *some* grounding of those thoughts in perception (not the particular sensory grounding which Mary doesn't yet have, but some grounding). There is no additional requirement that the agent be able to re-engineer, at will, the mechanisms governing all these reason respecting transitions, and this is what RoboDennett would have to do in order to use his simulated V4 to put himself into the functional state of knowing what it is like. On the account presented in Section VIII, you know what it is like to see red only when you possess the grounded concept '*red_as_experienced*'. That concept exists only when the relevant linkage between low and high level processing — or something functionally isomorphic to it — has been created. To get this grounding other than by low-level stimulation of the kind which normally engenders colour experience, an agent would need to re-engineer its cognitive architecture using abilities which go beyond those *required* by the knowledge argument. Lacking this low level grounding, RoboDennett simply wouldn't have this grounded concept — with its concomitant behavioural and affective results — even though he knows *exactly* what these results would be, if he did have the grounded concept in question.

[14] Certainly speed of processing *will* eventually be a problem if RoboDennett tries to generate a simulation of himself generating a simulation of himself… etc. But I think we should allow that RoboDennett only needs to go one level deep, and that he could unpick the differences in state due to the fact that he was running a simulation and the simulation wasn't, from those differences due to the fact that the simulation had experienced colour, and he hadn't.

Now as RoboDennett would not know what it is like, even while he runs all these incredibly complicated simulations, we are entitled to ask what it *would* be like for him to run them. I submit that it would be like nothing so much as it would be like thinking very hard! As we have said, the result of all that thinking very hard would be that RoboDennett would know exactly what he should say and how he would react if he had seen colour. So now we must address one final question: why can't RoboDennett simply speak and react as he knows he should?

XII: What Heterophenomenology Requires

Dennett has frequently, eloquently and correctly argued that a difference that makes no difference *is* no difference (Dennett, 1991; 1995; 2004).Take Dennett's position on philosophical zombies, for instance. A zombie is a creature which responds to any stimulus which experimenters present to it in exactly the way we would. Thus a zombie may well decide to stand there all day saying things like 'Of course I have qualia! Why won't you believe me, dammit?', not just in the manner of an over-complex lookup table, but in *all* the same ways and on all the same occasions we would, tested and untested.

Dennett's response to this thought experiment — the correct response — is to believe the zombie. Of course it has qualia.[15] To think otherwise is to make a fundamental mistake about the nature of introspection, a mistake which leaves each of us as the proud owners of our own epiphenomenal qualia.

So we need to make very sure that RoboDennett is not an unintended zombie. To sustain the claim that RoboDennett does not know what it is like, we need to demonstrate that he *cannot* behave exactly like a creature which does know what it is like.

We can demonstrate this by first noting that personal level behaviour does not consist simply in verbal (or other types cf. Marcel, 1993; Cowey and Stoerig, 1995) of report. There are additionally many things that we, as agents, do, over which we have no conscious, voluntary control. We sneeze in response to dust; we blink to protect our eyes, and duck to protect our bodies, from looming stimuli; we have certain innate, low level reactions to sound and, the case in point, to colour (Humphrey and Keeble, 1978).

If it is possible to build an agent who knows as much as Mary, but with our kind of hierarchical architecture, then these behavioural differences would remain. The very simplest example is speed of response: non-consciously mediated responses are simply faster (Marcel, 1993; Merikle, Smilek *et al*., 2001) than consciously mediated responses. Because of this, however much RoboDennett knows about how he should have reacted to any given coloured stimulus which he sees, he will be too late to *actually* react as fast as if the reaction had genuinely been mediated by lower level processes. This is a *bona fide* behavioural difference, and one which RoboDennett cannot overcome.

[15] To more accurately reflect Dennett's position, I should say: 'of course it is exactly as justified in claiming to have qualia as we are.'

There are also behavioural differences in kind, not just in speed, of response. Take the example of the heightened state of alertness in rhesus monkeys in response to red light reported by Humphrey and Keeble (1978). This change in behavioural pattern is mediated by an extremely complex set of biochemical changes, one which we very probably cannot create by any chain of conscious thought; crucially, though, whether or not we actually can do this, it is entirely reasonable to suggest that there is no logical or physical *entailment* from the ability to understand what such changes consist in to the ability to initiate such changes by any act of conscious will. Again, therefore, RoboDennett would lack these abilities, and simply would not be able to *behave* like a creature who had undergone the low-level changes which would occur in him after exposure to colour.

These low level abilities are a crucial part of what Mary gains, when she learns what it is like. She is said to know what it is like precisely because her more abstract concept, '*red_as_experienced*', is supported and enabled by the very systems which mediate faster, less abstract responses to red. A creature which really knows what it is like must really behave as if its low level systems have been exposed to colour, and it must also reason about colour, as experienced, in a way which is supported by those low level systems (with consequent two-way effects, from reasoning to low level responses and *vice versa).*

All of this RoboDennett would lack, despite his perfect knowledge of what he lacks. This will result in personal level behavioural differences, which he cannot overcome, between RoboDennett and an agent which does know what it is like.

Knowing as much as Mary does — knowing *exactly* what these low level behavioural differences consist in — does not entail the ability to simulate these behavioural differences perfectly (it *admits of* such an ability, but it *does not* entail it), thus RoboDennett, who can only do what he must be able to do in virtue of his knowledge, does not know what it is like, even on a strictly heterophenomenological account.

XIII: Conclusion

This paper began with a brief review of the current status of the major physicalist responses to the knowledge argument. I suggested that there is far more which unites these responses than there is which divides them. Recently Jackson himself has joined this coalition.

As has been pointed out elsewhere, Dennett seems to be ploughing a lone furrow on this argument. Now it is unwise to write off Dennett's lone furrows. They tend to be at worst well argued and informative, and at best — and often — correct despite the nay-sayers. In this instance, however, I believe we can marshal strong arguments for the former outcome.

Dennett's recent, clear statement of his position on the knowledge argument again emphasises what we already knew, that he requires an account of knowing what it is like which is multiply realizable, and which is fully compatible with heterophenomenology. Dennett has consistently taken it to be the case the these

two requirements (especially, perhaps, the second) must result in a non-qualia-realist account of subjective experience. The positive work of this paper has been to take Dennett's framework — and these two pre-conditions — seriously, and then to argue for precisely the opposite conclusion: that there exists a physicalist, functionalist, heterophenomenological, but essentially qualia-realist account of subjective experience.

In order to reach this conclusion we have had to consider three key points:

(1) What exactly counts as cheating in the knowledge argument? We have made explicit what is probably the only workable criterion: that Mary (or any surrogate agent whom we wish to discuss) must be allowed only those abilities which are necessitated by (a) her knowing and understanding as much as she does about the facts of colour vision and (b) her potential to come to know what it is like in the way in which we do.
(2) To work successfully with the second part of this definition, we needed to nail our flag to the mast, and to present a working analysis of the state of knowing what it is like. I have suggested that only a few steps beyond the original responses of Churchland and Lewis lies an account in which the state of knowing what it is like is analysed as that state in which one's more abstract reasoning about colour as experienced is actively supported by lower level processing dedicated to categorising and reacting to colour.
(3) Armed with the two analyses above we have looked carefully at what physicalism and heterophenomenology require. It turns out that on *any* physicalist account an agent who knows as much as Mary can do an awful lot, just by thinking about it. In particular, she can generate a *bona fide* state of knowing what it is like, without exceeding the abilities granted to her *ex hypothesi* by the knowledge argument. But this state of knowing what it is like is a state of an internal model of herself (which she can, necessarily, generate) and it is not, without further work, a state which she, the modeller, is in. The further steps which Mary must take in order to put herself into this state — the state which she knows so much about — are steps which require abilities beyond those granted by the knowledge argument.

Crucially, in addition to presenting arguments about Mary's internal state, we have also looked at Mary's predicament purely from the outside. On a heterophenomenological account, Mary does not have to pass any kind of arbitrary, inner-directed test about what state she is in. All she has to do is to be able to choose to behave exactly as if she knew what it is like, if she so wishes. On a heterophenomenological account, that is enough for her to know what it is like.

And it is enough. But she cannot do it.

In us, in Mary, and in any system in which high-level knowledge is supported by lower-level processing, all the high-level knowledge you could wish for *about* lower-level processing remains insufficient *ipso facto* to cause an instantiation of that lower level processing, in the right causal relationship to the higher level processing.

Thus RoboMary is completely physical, but to come to know what it is like she has to cheat, under as rigorous a definition of cheating as you could wish for. Adapting Dennett's own framework, we have therefore defined another robot agent, RoboDennett, who knows as much as Mary, but who cannot cheat: he cannot use abilities beyond those granted by the premises of the knowledge argument. Thus RoboDennett, not RoboMary, is the correct intuition pump surrogate for Mary herself.

Detailed consideration of what might and must be true of such agents has led us to credit the Mary intuition, after all. RoboDennett, despite all his knowledge — and thus Mary herself, despite all her knowledge — still would not know what it is like, even on a strictly physicalist, functionalist and heterophenomenological account.

Acknowledgements

This paper is a significantly revised and clarified version of a paper with the same title which was awarded *Best Student Paper* at the *Toward a Science of Consciousness* conference, Tucson, AZ, 2004.

For input at various stages, I would like to offer my warmest thanks to Steve Torrance, Marco Giunti, Simon McGregor, Rowan Lovett, Ron Chrisley, Torin Alter, Dan Dennett, Dave Chalmers and an anonymous referee.

References

Adams, F. and K. Aizawa (2001), 'The bounds of cognition', *Philosophical Psychology*, **14** (1), pp. 43–64.

Alter, T. (1998), 'A limited defense of the knowledge argument', *Philosophical Studies*, **90**, pp. 35–56.

Alter, T. and S. Walter (2006), *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism* (New York: Oxford University Press).

Chalmers, D.J. (1994), 'On implementing a computation', *Minds and Machines*, **4** (4), pp. 391–402.

Chalmers, D.J. (2003), 'The matrix as metaphysics', *http://consc.net/papers/matrix.html* Accessed 2005 (7th July).

Chrisley, R. (1994), 'Why everything doesn't realize every computation', *Minds and Machines*, **4** (4), pp. 403–20.

Churchland, P.M. (1985), 'Reduction, qualia and the direct introspection of brain states', *Journal of Philosophy*, **82**, pp. 8–28.

Churchland, P.M. (1989), 'Knowing qualia: A reply to Jackson', *On The Contrary*, P.M. Churchland and P.S. Churchland (Cambridge, MA: MIT Press).

Churchland, P.M. (1998), 'Postscript to knowing qualia', *On the Contrary*, P.M. Churchland and P.S. Churchland (Cambridge MA: MIT Press).

Cowey, A. and P. Stoerig (1995), 'Blindsight in monkeys', *Nature*, **373**, pp. 247–9.

Dennett, D.C. (1988), 'Quining qualia', *Consciousness in Modern Science*, ed. A. Marcel and E. Bisiach (Oxford: Oxford University Press).

Dennett, D.C. (1991), *Consciousness Explained* (Boston, MA: Little, Brown & Co.).

Dennett, D.C. (1994), 'Get real', *Philosophical Topics*, **22** (1&2), pp. 505–68.

Dennett, D.C. (1995), 'The unimagined preposterousness of zombies: Commentary on T. Moody, O. Flanagan and T. Polger', *Journal of Consciousness Studies*, 2 (4), pp. 322–6.

Dennett, D.C. (2004), 'Consciousness: How much is that in real money?', *Oxford Companion to the Mind, 2nd Edition*, ed. R.L. Gregory (Oxford: Oxford University Press).

Dennett, D.C. (2005), 'What RoboMary knows', *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness*, D.C. Dennett (New York: Oxford University Press).
Dennett, D.C. (2006), 'What RoboMary knows', *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, ed. T. Alter and S. Walter (New York: Oxford University Press).
Fuster, J.M. (2004), 'Upper processing stages of the perception–action cycle', *Trends in Cognitive Sciences*, **8** (4), pp. 143–5.
Graham, G. and T. Horgan (2000), 'Mary Mary, quite contrary', *Philosophical Studies*, **99**, pp. 59–87.
Humphrey, N.K. and Keeble, G.R. (1978), 'Effects of red light and loud noise on the rate at which monkeys sample the sensory environment', *Perception*, **7**, pp. 343–8.
Jackson, F. (1982), 'Epiphenomenal qualia', *Philosophical Quarterly*, **32** (127), pp. 127–36.
Jackson, F. (1986), 'What Mary didn't know', *Journal of Philosophy*, **83** (5), pp. 291–5.
Jackson, F. (1998), 'Postscript on qualia', *Mind, Method, and Conditionals* (London: Routledge).
Jackson, F. (1998), 'Preface', *Mind, Method, and Conditionals* (London: Routledge).
Jackson, F. (2003), 'Mind and illusion', *Minds and Persons*, ed. A. O'Hear (Cambridge: Cambridge University Press).
Laureys, S., Faymonville, M.E. *et al*. (2004), 'Residual functioning in the vegetative state', *Life Sustaining Treatments and the Vegetative State, Rome 17-20 March, 2004*.
Lewis, D. (1980), 'Mad pain and Martian pain', *Readings in the Philosophy of Psychology: Volume I*., ed. N. Block (Cambridge, MA: Harvard University Press).
Lewis, D. (1983), 'Postscript to "Mad pain and Martian pain"', *Philosophical Papers: Volume I* (Oxford: Oxford University Press).
Marcel, A.J. (1993), 'Slippage in the unity of consciousness', *Ciba Foundation Symposium No. 174 Experimental and Theoretical Studies of Consciousness*, ed. G.R. Bock and J. Marsh (Chichester: John Wiley & Sons).
Merikle, P.M., Smilek, D. *et al*. (2001), 'Perception without awareness: Perspectives from cognitive psychology', *Cognition*, **79**, pp. 115–34.
Nemirow, L. (1980), 'Review of *Mortal Questions* by Thomas Nagel', *Philosophical Review*, **89**, pp. 473–77.
Nemirow, L. (1990), 'Physicalism and the cognitive role of acquaintance', *Mind and Cognition: A Reader*, ed. W.G. Lycan (Oxford: Blackwell).

Paper received May 2004; revised September 2005.