THE TROUBLE WITH MARY

BY

VICTORIA MCGEER

Abstract: Two arguments are famously held to support the conclusion that consciousness cannot be explained in purely physical or functional terms – hence, that physicalism is false: the modal argument and the knowledge argument. While anti-physicalists appeal to both arguments, this paper argues there is a methodological incoherence in jointly maintaining them: the modal argument supports the possibility of zombies; but the possibility of zombies undercuts the knowledge argument. At best, this leaves anti-physicalists in a considerably weakened rhetorical position. At worst, it shows that commonsense intuitions on which anti-physicalists rely mislead us about the true nature of conscious experience.

I.

Physicalism is the doctrine that there is nothing over and above the physical constituents of our world. All real phenomena, including in particular mental phenomena, are ultimately physical phenomena, explicable (perhaps non-reductively) in recognizably physicalistic terms. The exact form these explanations will take is much debated, but the current most favoured bet is that even the most recalcitrant features of conscious experience (e.g., what it's like to see a red tomato) will be analyzable in unproblematically *functional* terms – i.e., in terms of what the constituents of complex physical systems *do* – even if the appropriate level of functional explanation is one of fine-grained neurological activity (Dennett, 2001).

Physicalism as such defines a philosophical-scientific project. Against this project range a variety of anti-physicalistic nay-sayers, or "phenomenalists" as I will also call them. Minimally, phenomenalists contend that the functionalist principles governing current investigations into the nature of our complex cognitive capacities are inadequate to the task of

Pacific Philosophical Quarterly 84 (2003) 384–393 0279–0750/00/0100–0000

explaining the nature and purpose of conscious experience itself (Block, 1995; Block & Stalnaker, 1999; Braddon-Mitchell & Jackson, 1996; Chalmers, 1996; Levine, 1983; Levine, 1993; Nagel, 1974). Some take this to suggest a more radical epistemic conclusion that the link between mind and body can never be understood by minds such as ours, since we access the intrinsic and extrinsic properties of our own nature in such radically different ways (Chomsky, 1994; McGinn, 1999; Nagel, 1998). Others embrace the more radical metaphysical view that consciousness really is something over and above any physical phenomena that may be correlated with it (Chalmers, 1996; Eccles, 1994; Jackson, 1982; Jackson, 1986).[1]

While phenomenalists bring various considerations to bear in support of their anti-physicalist stance, two arguments are given pride of place – the modal argument and the knowledge argument. As we shall see, these arguments are interesting not just because of what each purports to establish, but also in how they allegedly fit together.

The modal argument turns on the intuition that certain things are conceivable and therefore possible which ought not to be possible if physicalism is true. The main intuition rehearsed under this argument is that zombies are conceivable, where zombies are exact physical and hence functional replicas of conscious human beings, save for the fact that they have no inner mental life. They are not *phenomenally* conscious at all (Block, 1990). If zombies are conceivable, so the argument goes, then they are possible. And if zombies are possible, then physicalism is false since physicalism is committed to the view that the physical and functional properties of conscious human beings – the properties present in zombies – are logically sufficient to ensure consciousness.[2]

The standard physicalist response to the modal argument is either to question the inference from conceivability to possibility or to question the claim that zombies really are conceivable. I won't elaborate on either of these responses directly, since they have been discussed extensively by others and are not particularly germane to the methodological point I want to emphasize.[3] Suffice to say that much of the heat generated by the modal argument may stem from the fact that it seems to buy its conclusion too easily. Indeed, some of the proponents of anti-physicalism (e.g. Frank Jackson) are inclined to call it an intuition rather than an argument proper. But, even as an intuition, it has what many consider to be some deeply puzzling consequences.

Take, for instance, the "paradox of phenomenal judgement" (Chalmers 1996, Chapter V). Since zombies are physical (and therefore functional) replicas of human beings, they make judgements just as we do about what they take to be their own conscious experiences – e.g. that they're having a red sensation now, that having a red sensation is *like this*, that having red sensations is a mysterious kind of thing, and so forth. Judgements,

after all, are just behavioural acts – as much a part of the physical world as verbal reports. Consequently, the neurophysiological processes that subserve cognition are sufficient to explain the occurrence of phenomenal judgements in the zombies' case, just as they are sufficient to explain the occurrence of these judgements in our own case. The only difference between zombies and us, according to anti-physicalists, is that all the zombies' phenomenal judgements are wrong, and wrong because they have no experiences to be right about. So, it seems that consciousness in the phenomenal sense is explanatorily and causally irrelevant to the making of phenomenal judgements, though not to their justification or truth-value. Chalmers himself calls this paradoxical situation "delightful and disturbing" but "not obviously fatal to the nonreductive position" (Chalmers, 1996, p. 181). What makes him so confident?

Enter the knowledge argument: Jackson's alleged demonstration that there are non-physical facts known about experience over and above all the physical facts that characterize our universe. The knowledge argument proceeds by imagining Mary, a super-scientist confined to a black-and-white room, who comes to know everything there is to know about the physical nature of the world. In particular, she knows everything there is to know about the physical and functional nature – the "structure and dynamics" – of human visual processing under all conceivable circumstances, including retinal stimulation provided by a red tomato. Nevertheless, upon being released from her black-and-white room and seeing a red tomato for the first time herself, something new happens to her that seems appropriately described as coming to learn something new – namely, what it is like to have a red tomato experience. Various commentators have protested that Mary does not learn any new facts by virtue of having this new experience, but only a new skill or something of the kind (Lewis, 1990; Nemirow, 1990). But given that she also learns something new about the mental life of *others* – presumptively, what their colour experiences are like – such knowledge would be hard to analyze in a non-propositional way. As Jackson forcefully puts it:

> The trouble for physicalism is that, after Mary sees her first ripe tomato, . . . she will realize that there was, all the time she was carrying out her laborious investigations into the neurophysiologies of others and into the functional roles of their internal states, something about these people that she was quite unaware of. All along their [red visual] experiences . . . had a feature conspicuous to them but until now hidden from her. But 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. But it is a fact about them. That is the trouble for physicalism (Jackson, 1986).

So it seems that physicalists have two arguments to combat. Worse, as Chalmers explains, these are two arguments whose persuasive sum is importantly greater than their parts:

We have seen that the modal argument (the argument from logical possibility) and the knowledge argument are two sides of the same coin. I think that in principle each succeeds on its own, but in practice they work best in tandem. Taking the knowledge argument alone: most materialists find it hard to deny that Mary gains knowledge about the world, but often deny the step from there to the failure of materialism. Taking the modal argument alone: most materialists find it hard to deny the argument from the conceivability of zombies or inverted spectra to the failure of materialism, but often deny the premise. But taking the two together, the modal argument buttresses the knowledge argument where help is needed, and vice versa. In perhaps the most powerful combination of the two arguments, we can use the knowledge argument to compellingly establish the failure of logical supervenience, and the modal argument to compellingly make the step from that failure to the falsity of materialism (Chalmers, 1996, pp. 145–6).

Chalmers' idea seems to be this. The knowledge argument provides data – namely, Mary's seeming to learn something new – that is physicalistically hard to explain. The modal argument – the intuition that zombies are possible and, more generally, that consciousness is such that zombies (and other alien creatures) are possible – directs us to a way of explaining this data. The knowledge argument on its own gives us data without any clear inferential destination. The modal argument on its own gives us a potential inferential destination without any substantive data to take us there. The knowledge argument without the modal argument is blind in the sense of destination-deprived. The modal argument without the knowledge argument is empty in the sense of data-deprived.

Jackson evidently concurs with Chalmers' reading of the relationship between these two arguments. Not only does Chalmers thank him for discussing the point in a footnote to the passage cited above, Jackson expresses a similar view in his initial presentation of them: "*qua* protagonists of the Knowledge argument we may well accept the modal intuition in question; but this will be a *consequence* of our already having an argument to the conclusion that qualia are left out of the physicalist story, not our ground for that conclusion" (Jackson, 1982, p. 472).

How best to respond to this tandem attack on physicalism? In martial arts, a well-known strategy for bringing the opposition down is to use their own weight against them. So instead of defending physicalism by disputing these arguments individually, it may just be simpler to do what these phenomenalists suggest and consider their arguments together – that is, in genuine combination. It turns out, as we shall now see, that far from being mutually supporting, the modal argument and the knowledge argument end up in tension with one another.

II.

Assume that the modal argument is sound and that zombies are therefore a genuine possibility. Now imagine Mary's zombie duplicate confined to

her black and white room and learning all that there is to know about the neurophysiology of colour experience. Upon her release, Zombie Mary is confronted with the proverbial red tomato, and so sees something red for the first time herself. Presumably this means her visual system is now affected in the way she has learned it would be affected under such circumstances and, indeed, the way the visual systems of other normally sighted people would presumably be affected. But does she learn some new (non-physical) fact about the world? As Mary's zombie duplicate, she certainly reports having a new visual experience. She judges herself to have that experience. And she judges that others' visual experiences had all along a feature conspicuous to them that was hidden from her until now. But, as Chalmers explicitly agrees, she must be wrong about all this since she's only a zombie.[4] And so it must be that she does not learn some new fact about the world after all. Consequently, anti-physicalists must agree that, in her world, Zombie Mary's seeming to learn something new is not a reliable indicator that something has been left out of a wholly physicalistic account of her transformation.

This is not an easy conclusion for anti-physicalists to embrace. It means that any data the knowledge argument generates cannot be interpreted consistently across possible worlds. Depending on context, it must be taken either to an anti-physicalist conclusion (Mary learns a non-physical fact) or to a physicalist conclusion (Zombie Mary does not learn a non-physical fact). This move saves anti-physicalists from strict inconsistency. But it is purchased at the price of a methodological instability. For consider what their position now requires. According to Chalmers' insistence that these two arguments work in tandem, the knowledge argument provides clear data, and the modal argument provides a clear destination to which those data ought to take us. But the response to Zombie Mary – that she doesn't *really* learn any new fact – works only on the assumption that the modal argument licenses us sometimes to neglect the data provided by the knowledge argument. The tandem claim presumes that the modal argument gives the knowledge argument the power to carry us to an anti-physicalistic conclusion; but the response to Zombie Mary supposes that the modal argument can be invoked to explain why in this particular case the knowledge argument has no such power. This is a methodologically untenable position for anti-physicalists to occupy. It amounts to invoking one and the same factor, now to support the inference to a certain conclusion (that some new fact is really learned), now to explain why that inference is not compelling.

In light of this, it is not clear where Chalmers and his fellow travellers should go. But it is clear that the very possibility of zombies undercuts the anti-physicalists' independent appeal to the knowledge argument. With that possibility alive, the knowledge argument generates the data that Zombie Mary as well as Mary seems to learn something new on seeing the red

tomato. In holding that an interpretation of this data must be modally relativized, anti-physicalists are not only forced into a methodological incoherence whereby they give weight to the knowledge argument in one context and no weight to it in the other. They must also concede that the knowledge argument does no real work in keeping the anti-physicalist barge afloat. Everything now rests on the modal intuition which dictates how the knowledge argument is to be understood. This is a considerably weaker argumentative position for anti-physicalists to find themselves in, even by their own lights.

III.

We have seen that the modal intuition has the unintended consequence of turning Chalmers' tandem claim on its head: Instead of using "the knowledge argument to compellingly establish the failure of logical supervenience, and the modal argument to compellingly make the step from that failure to the failure of materialism" (*op. cit.*, p. 146), anti-physicalists are committed to using the modal intuition to back the knowledge argument in establishing the failure of logical supervenience, but only in those worlds where materialism is false. This is bad enough from a rhetorical point of view; but can the knowledge argument really succeed even in this limited task? There are reasons to doubt it.

Return for a moment to Zombie Mary. If anti-physicalists concede that she is a possible being (and how can they not?), it seems they owe us some account of the change Zombie Mary judges herself to have undergone in being released from her black and white room – an account that does the job of explaining what she takes to be revelatory in this transformation, without appealing to any new knowledge of non-physical facts. Whatever its details, it seems reasonable to concede that it would have to be something like the kind of account certain functionalists hoped would succeed in explaining what happens to Mary herself when she has her novel colour experiences (In fact, this is just what Chalmers maintains in Chalmers (2000 (web)/2002).

But now anti-physicalists are caught in a bind. For consider just what this concession entails. It is now agreed that Zombie Mary undergoes a transformation in her discriminatory abilities of a kind materialists would be happy to embrace. As a result, it's also agreed that she judges incorrectly that she learns some (non-physical) fact about the nature of red tomato experiences – a fact, which despite her exhaustive knowledge of all things physical, she did not yet know. In other words, it is agreed that undergoing this transformation is sufficient to induce in Zombie Mary a phenomenal illusion – i.e., an illusion about something she now calls a physically inexplicable property of (her own and others') discriminating subjectivity. But if Zombie Mary's physical transformation is sufficient to

induce such an illusion in her, it is surely sufficient to do so in Mary as well – they are, after all, physical replicas of one another. So, assuming we have an adequate explanation for why Zombie Mary seems to learn a new fact on leaving the black and white room, it must be agreed that we have an adequate explanation for why Mary herself seems to learn a new fact under the same circumstances. The challenge for anti-physicalists is now to say why Mary's seeming to learn a new fact on seeing the red tomato involves her really learning a new fact, without begging the question the knowledge argument is supposed to establish – namely, that her experienced transformation must consist in her learning such facts.

Anti-physicalists may insist that we can know such is the case with Mary by simply imagining ourselves in her place, as "fully conscious" creatures with "real phenomenal properties" that are instantiated the moment we spy a red tomato. But why should we be less susceptible to illusion than Mary herself? For all we know, the "real phenomenal properties" we think we instantiate in our own experience – our very own qualia – are nothing more than the illusory reification of complex discriminatory capacities in us as well. Hence, our transformation on leaving the black and white room would no more involve our learning a new fact than would Mary's or Zombie Mary's – though, of course, we 'might think that it does' (Chalmers, 2000 (web)/2002). In general, once the capacity for such illusions has been granted to creatures that have no consciousness in the preferred sense, the knowledge argument ceases to be rationally compelling. For we may all be such creatures in the end – and that possibility alone constrains us to admit that any creature's judgement about phenomenal properties can be given a deflationary physicalistic explanation in terms of discriminatory abilities that, for all anyone can tell, is perfectly adequate to the case at hand.

There is one final curiosity that the possibility of Zombie Mary leaves unaddressed. I have claimed that anti-physicalists are compelled to concede that, in so far as Zombie Mary suffers from a phenomenal illusion when she leaves the black and white room, it can be adequately explained in terms of her acquiring new discriminatory abilities. But this may not be a viable option after all. Consider for a moment the assumption of Zombie Mary's physical omniscience – i.e., the assumption that she knows all the physical facts there are to know. Does this imply that, in the domain of physical knowledge, she is free from ignorance *simpliciter*, or is she free from error as well? If she is free from ignorance *simpliciter*, then the argument of the preceding paragraph stands: Zombie Mary can be mistaken about the nature of her own physical transformation – suffer a phenomenal illusion – since this is consistent with her knowledge of all things physical. But if her physical omniscience implies freedom from error about physical things as well as freedom from ignorance, then a still more dramatic result follows.[5] Zombie Mary as she figures in the zombie knowledge

argument will not be a possible being, since no such physically omniscient creature could be deceived in the requisite way.

This in turn will have hugely dramatic consequences for the anti-physicalist position. If Mary herself is a possible being, then, according to the anti-physicalist recipe for constructing zombies, Zombie Mary ought to be possible too. Hence, by modus tollens, we reach the happy conclusion that Mary herself is not a possible being, where Mary, recall, is a creature that knows all the physical facts there are to know, but still learns some hitherto unknown fact when first she sees a red tomato.[6] Physicalists, of course, will be delighted with this result: It gives them good reason to insist that, however commonsensical anti-physicalist intuitions may seem, they do not hang together well in the final analysis, so cannot make a reliable guide for exploring the true nature of conscious experience.[7]

Philosophy Program,
RSSS, Australian National University
and
Department of Philosophy,
Simon Fraser University

NOTES

[1] More recently, Jackson has refused to draw this, or any, anti-physicalist conclusion from the knowledge argument. As he says himself, "I now think that the puzzle posed by the knowledge argument is to explain why we have such a strong intuition that Mary learns something about how things are that outruns what can [be] deduced from the physical account of how things are" (Jackson, 1998, pp. 77–78).

[2] Other versions of the modal argument concern the conceivability (and, therefore, possibility) of physical replicas that have alien or inverted experiences. All are inconsistent with physicalism, since they propose mental phenomena that vary independently of a creature's physical constitution. For a vivid presentation of these arguments, see Chalmers (1996).

[3] Among these discussions is a particularly interesting paper by Katalin Balog (Balog, 1999). She attacks the modal argument, and in particular the inference from conceivability to possibility, by means of a thought-experiment that has similar features to the one I offer in Section II below. Our arguments against anti-physicalism are different since I do not question the inference from conceivability to possibility, but I am sympathetic both to Balog's argumentative strategy and to her conclusion.

[4] For Chalmers' discussion of how to interpret Zombie Mary's phenomenal judgements, see (Chalmers, 2000(Web)/2002).

[5] There is reason to think that Zombie Mary must be free from error – at least of the required sort – if we assume that she is free from ignorance in the physical domain. For if Zombie Mary is subject to the phenomenal illusion in question, then it seems that there is some physical fact that she fails to recognize as such – namely, that her "phenomenal" transformation is purely physical in character. Why count this as a physical fact? Consider the following analogous case: If it's a physical fact that water is H_2O and also a physical fact that H_2O is nothing but a physical substance, then surely it is a physical fact that water

is nothing but a physical substance. By analogy, if it's a physical fact that Zombie Mary's "phenomenal" transformation is a particular neurological process, and it's a physical fact that this neurological process is nothing but a physical process, then surely it's a physical fact that Zombie Mary's "phenomenal" transformation is nothing but a physical process. But if Zombie Mary fails to recognize this fact about herself, then there is some physical fact of which she is ignorant, contrary to the assumption that she knows all the physical facts there are to know. So it seems more consistent to maintain that omniscience qua freedom from ignorance in the domain of physical facts implies freedom from error in that domain as well – or at least freedom from error of the sort required for her supposed phenomenal illusion.

[6] This leaves open two possibilities: (1) Mary does not know all the physical facts there are to know before she leaves the black and white room – i.e., some physical facts can only be learned through direct experience; or (2) Mary *does* know all the physical facts before leaving the room (she is an imaginative construct, after all), but then (*a fortiori*) she also knows what it's like to see a red tomato (Dennett, 1991).

[7] My thanks to Dan Dennett, Frank Jackson, Philip Pettit, David Rosenthal, and Daniel Stoljar for helpful comments on an earlier draft. Thanks also to an anonymous reviewer for *Pacific Philosophical Quarterly* for suggesting an important amendment to my final argument.

REFERENCES

Balog, K. (1999). "Conceivability, Possibility, and the Mind-Body Problem," *Philosophical Review* 108(4), pp. 497–528.

Block, N. (1990). "Consciousness and Accessibility," *Behavioral and Brain Sciences* 13(4), pp. 596–598.

Block, N. (1995). "On a Confusion about a Function of Consciousness," *Behavioural and Brain Sciences* 18, pp. 227–47.

Block, N., & Stalnaker, R. (1999). "Conceptual Analysis, Dualism and the Explanatory Gap," *Philosophical Review* 108(1), pp. 1–46.

Braddon-Mitchell, D., & Jackson, F. (1996). *Philosophy of Mind and Cognition*. Oxford: Blackwell.

Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.

Chalmers, D. (2000(Web)/2002). "The Content and Epistemology of Phenomenal Belief," in A. Jokic & Q. Smith (eds.), *Aspects of Consciousness*. Oxford: O.U.P.

Chomsky, N. (1994). "Naturalism and Dualism in the Study of Mind and Language," *International Journal of Philosophical Studies* 2, pp. 181–209.

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

Dennett, D. (2001). "The Zombic Hunch: extinction of an intuition?" *Philosophy* 48(Supp), pp. 27–43.

Eccles, J. (1994). *How the Self Controls its Brain*. Berlin, New York: Springer-Verlag.

Jackson, F. (1982). "Epiphenomenal Qualia," *Philosophical Quarterly* 32, pp. 127–36.

Jackson, F. (1986). "What Mary Didn't Know," *Journal of Philosophy* 83, pp. 291–5.

Jackson, F. (1998). "Postscript on Qualia," *Mind, Method and Conditionals: Selected Essays* London: Routledge, pp. 76–79.

Levine, J. (1983). "Materialism and Qualia: The Explanatory Gap," *Pacific Philosophical Quarterly* 64, pp. 354–61.

Levine, J. (1993). "On Leaving Out What its Like," in M. Davies & G. Humphreys (eds.), *Consciousness: Psychological and Philosophical Essays*. Oxford: Basil Blackwell.

Lewis, D. (1990). "What Experience Teaches," in W. G. Lycan (ed.), *Mind and Cognition: A Reader*. Cambridge, Mass: Blackwell, pp. 499–519.

McGinn, C. (1999). *The Mysterious Flame: Conscious Minds in a Material World*. New York: Basic Books.

Nagel, T. (1974). "What is it like to be a bat?" *Philosophical Review* 83, pp. 435–450.

Nagel, T. (1998). "Conceiving the Impossible and the Mind-Body Problem," *Philosophy* 73, pp. 337–52.

Nemirow, L. (1990). "Physicalism and the Cognitive Role of Acquaintance," in W. Lycan (ed.), *Mind and Cognition: A Reader*. Oxford: Blackwell, pp. 490–99.