Prolegomena 7 (1) 2008: 37–59

Mary's Scientific Knowledge*

LUCA MALATESTI
University of Hull – Philosophy, Hull HU6 7RX United Kingdom
l.malatesti@hull.ac.uk

ORIGINAL SCIENTIFIC ARTICLE / RECEIVED: 26–01–2007 ACCEPTED: 14–10–2007

ABSTRACT: Frank Jackson's knowledge argument (KA) aims to prove, by means of a thought experiment concerning the hypothetical scientist Mary, that conscious experiences have non-physical properties, called *qualia*. Mary has complete scientific knowledge of colours and colour vision without having had any colour experience. The central intuition in the KA is that, by seeing colours, Mary will learn *what it is like* to have colour experiences. Therefore, her scientific knowledge is incomplete, and conscious experiences have *qualia*. In this paper I consider an objection to the KA raised by Daniel Dennett. He maintains that the KA is vitiated by Jackson's account of Mary's scientific knowledge. While endorsing this criticism, I will defend the plausibility and relevance of the *type* of strategy involved in the KA by offering an account of Mary's scientific knowledge. This account involves formulating a reasonable and not immediately false version of the physicalist thesis with regard to colour experiences. Whether this version of the KA is successful against this type of physicalism is not investigated here.

KEYWORDS: Conscious experience, experiences of colours, knowledge argument, physicalism, *qualia*.

1. Introduction

Frank Jackson's knowledge argument is a famous objection to physicalism.[1] The most discussed version of this argument relies on a thought experiment concerning Mary, a vision scientist who has complete scientific knowledge of colours and colour vision, but has never had colour experi-

* I have presented versions of this paper in research seminars at the University of Siena (Italy) and the University of Rijeka (Croatia). I would like to thank the participants for their objections and suggestions. In addition, I am very grateful to the anonymous referee of this journal for criticisms that prompted substantial revisions.

[1] Jackson (1982) and Jackson (1986). However, Jackson has since renounced this; see Jackson (1998) and Jackson (2004).

ences. In fact, Mary has been held captive in a monochromatic environment, where she acquired her scientific knowledge only by seeing white, black and shades of grey. According to Jackson, when Mary is released from this environment and sees coloured objects, she acquires new knowledge that is not to be found among her complete scientific knowledge. He concludes that there are facts concerning colour experiences that scientific knowledge can neither describe nor explain. Specifically, these facts involve the occurrence of certain non-physical properties of experiences that he calls *qualia*.

I will argue that Jackson's knowledge argument requires major revision. In fact, Jackson does not offer a precise characterisation of Mary's scientific knowledge. As Daniel Dennett has shown, this lack of precision authorises legitimate interpretations under which his reasoning is unintelligible. Moreover, Jackson's account of Mary's scientific knowledge leaves open possibilities that could undermine the soundness of the knowledge argument.[2]

Nevertheless, I will argue that the *type* of strategy involved in Jackson's argument might still be viable. The central problem to be addressed here is that of striking the right balance between the intelligibility of what Mary knows before her release and the strength of the knowledge argument's conclusion. Jackson's argument, which is aimed at a highly abstract version of physicalism, is intended to show that there are non-physical facts and properties.

This paper offers a version of the knowledge argument that constitutes a substantive, interesting challenge to contemporary accounts of conscious colour experience. This version is directed against a "modest", well-defined, and reasonable formulation of an idea shared by many physicalists, namely, that science can accommodate colour experiences. While articulating and defending this formulation of the knowledge argument, I will not evaluate the soundness of such reasoning.

2. Knowledge Beyond Our Understanding

The knowledge argument can be formulated as follows:

(1) Mary, before her release, has complete scientific knowledge of the facts concerning colours and colour vision, without having had any conscious experiences of colours.

(2) By seeing a coloured object after her release, Mary acquires new knowledge about colour experiences.

[2] Dennett (1991: 399–403).

Therefore:

(3) There are facts that are not physical.

Let us consider the two premises of the argument. The second concerns knowledge that Mary allegedly acquires by having chromatic colour experiences. Jackson claims that this knowledge involves the occurrence of *qualia*. By the term *qualia* he wishes to refer to "certain features of the bodily sensations especially, but also of certain perceptual experiences" (Jackson, 1982: 469). According to Jackson, this suggests that, by seeing coloured objects, Mary comes to know facts about her colour experiences that involve the occurrence of non-physical *qualia*. For instance, when Mary sees a red object, she will learn that her experience of a red object has a certain feature. This feature is a *quale* that she did not know about before her release. I will refer to such knowledge as *knowledge of what it is like to have a colour experience*. Many have focused critical attention on this second premise of the knowledge argument.[3] However, the other premise likewise requires careful scrutiny.

The first premise of the knowledge argument involves a characterisation of Mary's scientific knowledge. This scientific knowledge should comprise the kinds of facts that, according to the physicalist, exhaust all there is to know about colour experience. Jackson maintains that he does not need a precise characterisation of physical facts or, in his words, "'physical information' and the correlative notion of physical property, process, and so on".[4] Instead, he assumes that it is enough for his purposes to say of Mary that "[s]he knows all the physical facts about us and our environment, in a wide sense of 'physical' which includes everything in completed physics, chemistry and neurophysiology" (Jackson 1986: 567). Additionally, he claims that Mary knows facts concerning the "functional roles" played by states of the nervous system.[5] This characterisation is, however, not satisfactory.

In the knowledge argument, the assumption that Mary, by seeing a coloured object, comes to know something that she did not know before her release plays a central role. The alleged novelty of this knowledge and, supposedly, our understanding of its nature should stem from the contrast with what she knew before her release. However, as Daniel Dennett has persuasively pointed out, we cannot contrast Mary's scientific knowledge with her knowledge of what it is like to have a colour experience.[6] In fact, grasping her scientific knowledge is a task "so preposterously immense,

[3] For an overview of this debate, see Nida-Rümelin (2002).
[4] Jackson (1982: 469).
[5] Jackson (1986: 567).
[6] A similar observation is offered by Churchland (1986).

you can't even try. The crucial premise is that 'She has all the *physical* information.' That's not readily imaginable […]" (Dennett 1991: 399). Moreover, the inadequate characterisation of Mary's scientific knowledge allows Dennett to offer a thought experiment to counter Jackson's argument.

Dennett assumes that if, before her release, Mary lacked knowledge of certain features of colour experiences, then upon her release she will be unable, when looking at a blue banana, to recognise that it has the wrong colour. He then tells the following story:

> And so one day, Mary's captors decided it was time for her to see colors. As a trick they prepared a bright blue banana to present as her first color experience ever. Mary took one look at it and said 'Hey! You tried to trick me! Bananas are yellow, but this one is blue!' Her captors were dumbfounded. How did she do it? 'Simple,' she replied. 'You have to remember that I know everything – absolutely everything – that could ever be known about the causes and effects of color vision. So of course before you brought the banana in, I had already written down, in exquisite detail, exactly what physical impression a yellow object or a blue object […] would make on my nervous system. So I already knew exactly what thoughts I would have (because, after all, the "mere disposition" to think about this or that is not one of your famous *qualia*, is it?). I was not in the slightest surprised by my experience of blue […] I realize that it is hard for you to imagine that I could know so much about my reactive dispositions that the way blue affected me came as no surprise. Of course it's hard for you to imagine. It's hard for anyone to imagine the consequences of someone knowing absolutely everything about anything!' (Dennett 1991: 399–400)

Some have maintained that Dennett uses this story to deny that Mary learns anything upon her release.[7] However, his aim here appears to be more modest: "My point is not that my way of telling the rest of the story proves that Mary *doesn't* learn anything, but that the usual way of imagining the story does not prove that she does. It doesn't prove anything […]" (Dennett 1991: 400). Thus Dennett is not arguing that Mary, by seeing coloured objects, does not undergo an epistemic progress. Rather, he tries to disqualify the intuition supporting the conclusion that she might learn something. There are also good reasons for thinking that this is the only conclusion we can draw from Dennett's story.

Dennett wants us to imagine a situation in which Mary does not learn anything new by having conscious experiences. He suggests that Mary's

[7] See Alter (1999) and Chalmers (1996: 145). This reading is probably suggested by the fact that Dennett has argued in other places for eliminating *qualia*; see Dennett (1988).

complete knowledge might consist of laws stating causal correlations between types of experiences and types of thoughts. In particular, Mary knows that if a subject is presented with a blue banana, she will have a thought expressible as: "This is blue". Thus, by seeing the blue banana, Mary comes to know that people who have this kind of experience think: "This is blue". Therefore, we can have intuitions about Mary's case that do not support the knowledge argument. However, does Dennett's story add anything to the observation that we cannot grasp Mary's scientific knowledge?

It appears that Dennett's account of Mary's case exploits the same lack of understanding of her scientific knowledge which he denounces in Jackson's knowledge argument. Indeed, let us assume, for the sake of argument, that there can be scientific laws connecting scientifically described experiences with thoughts.[8] Also, although Dennett does not explain how this is possible, we can concede that Mary's ability to recognise the *colour* of the object she sees enables her to know what it is like to have a certain *colour experience*.

Nevertheless, in order to judge whether Mary passes the blue banana test, we need a more substantial understanding of her scientific knowledge. A central assumption in Dennett's story is that Mary, by just looking at the banana, is able to apply the law correlating the conscious experience she is having with the thoughts usually caused by this type of experience.[9] As Dennett puts it: "Mary took one look at it and said 'Hey! You tried to trick me!'". This requires that Mary can recognise the colour experience involved in seeing the blue banana as a certain "physical impression" on her nervous system. However, how does she acquire such recognitional capacity? It seems that we are not in a position to say. On the other hand, it could be maintained that a relevant factor in our conclusion that Mary was not fooled might lie in on our understanding of her notion of "physical impression". This decision cannot, however, be based on Dennett's sketchy model of Mary's scientific knowledge. It seems that a more substantial account of her scientific knowledge is needed.

To sum up, it appears that Dennett's response makes two legitimate demands on supporters of the knowledge argument. First, they should provide an intelligible account of what Mary knows before seeing colours. Secondly, they have to prove that Mary cannot acquire the recognitional capacities involved in Dennett's story just by possessing this scientific

[8] Some have argued that thoughts cannot enter into any nomological correlation; see, for example, the very influential Davidson (1970).

[9] Howard Robinson shows clearly how this assumption is implicit in Dennett's story (Robinson, 1993: 175).

knowledge. In the following sections, I will argue that these two demands can be satisfied.

3. Defining the Physical

Our task is to offer an intelligible characterisation of Mary's scientific knowledge. In addressing this challenge, we can begin with the plausible assumption that the knowledge argument attacks a version of physicalism which has a *theory-based conception* of being physical. According to this conception, a certain entity is physical when a certain physical theory includes expressions that can refer to it. In fact, the knowledge argument requires that Mary knows completed physics, chemistry and neuroscience, and that if a fact escapes this knowledge, then it is non-physical.

A version of physicalism with a theory-based conception of being physical faces a dilemma.[10] On the first horn of this dilemma, if we define "being physical" by referring to some ideally complete future (or possible) physical theory, the notion of "physical" appears to lack content. In fact, we cannot predict what entities or laws an ideally complete future (or possible) physics might refer to.

On the second horn of the dilemma, if the definition of "physical" is based on current physical theory, then physicalism is false. In fact, if the physical is defined as whatever is described by current microphysics or other contemporary scientific doctrines, then new particles discovered by science in the future will not be physical, and thus the physicalist claim that everything is physical would be false. As seen in the previous section, Jackson's claim that Mary knows *all* the physical facts in completed physics and other sciences appears to expose his reasoning to the first horn of the dilemma.[11] On the other hand, if Mary's scientific knowledge is grounded in contemporary physics, the knowledge argument would be a redundant objection to a false doctrine.

A supporter of the knowledge argument might avoid these difficulties by offering an account of "being physical" which is alternative to the theory-based account. For example, by endorsing an *object-based conception* of the physical, physical entities might be defined as those of the same type as some entity taken to be paradigmatically physical.[12] The central idea in this account is that paradigmatically physical entities can be intro-

[10] On this problem with physicalism, see Hempel (1980), Crane and Mellor (1990) and Montero (1999).

[11] This reading is supported by Dennett (1991: 399–403) and Churchland (1986: 331–334). See also Mellor (1993).

[12] See Jackson (1998: 6–7).

duced without any reference to contemporary, future or possible science. Non-sentient ordinary objects such as tables and chairs can be taken as paradigmatically physical. Thus we might define physical properties and relations as those which are required in order to describe paradigmatically physical entities.

This object-based conception of the physical has its problems. The possibility of *panpsychism* threatens this account of "being physical".[13] Panpsychists believe that every entity has a mind. On this view, the paradigmatically physical objects referred to in the object-based account of the physical have a mind. Nevertheless, the object-based account requires the existence of non-sentient objects, and so proponents of this view must exclude the possibility of panpsychism. However, their theory does not seem to have the resources to specify further the nature of these non-sentient objects. It appears that they merely exclude the possibility of panpsychism by definition. However, although panpsychism may strike us as completely implausible, the possibility of it cannot be ruled out by stipulation alone. It seems we need some substantive account of the nature of ordinary objects in order to rule out that they have a mind.

The difficulty created by the possibility of panpsychism is an instance of a deeper problem with the object-based account. This account of the physical relies on the idea that we can have an ordinary understanding of what type of properties might figure in a complete account of objects such as chairs and tables. However, our ordinary understanding of these properties might turn out to be inadequate. It is enough to consider the image of reality provided by contemporary physics. The ultimate particles, properties and laws that this science invokes in providing an account of ordinary objects are very different from those we can contemplate in ordinary experience.[14] Thus a version of physicalism with an object-based account can even contradict physics, and endorsing such a theory requires facing the possibility that many of the basic assumptions of contemporary physics are false. Not many physicalists would be willing to develop an account of physical reality alternative to the one that contemporary physics provides.

The problems afflicting the theory-based and object-based conceptions of being physical might be avoided by denying one assumption that they share, namely, the idea that formulating physicalism requires a characterisation of physical entities.[15] Joseph Levine, for example, has recently argued that physical properties should be defined *per via negativa*

[13] Jackson (1998: 7).

[14] A criticism of this type can be found in Levine (2001: 20).

[15] This account is offered in Montero (1999). See also Levine (2001: 20–21).

in philosophical discussions of the mind-body problem. He contends that we have a clear enough grasp of mental properties. In particular, the mental is characterised by *phenomenal properties*, which specify what it is like to have mental states, and *representational properties*, which define the content of mental states. Thus he suggests that physicalism (or, in his words, "materialism") should be understood as the thesis that non-mental properties have ontological and explanatory priority over mental ones.[16]

This characterisation of the physical *per via negativa* should be rejected. In fact, it undermines one very good reason for endorsing physicalism. Physicalists place confidence in the explanatory power of scientific knowledge, by virtue of current developments in contemporary science. In particular, many physicalists have been impressed by results in biology and neuroscience that have explained many aspects of both normal and pathological human behaviour.[17] Thus it is central to the physicalist project to offer a conception of the mind that is not only consistent with contemporary science, but could also aid scientific progress.[18]

Clearly, justifying physicalism by referring to current scientific practice must involve a theory-based conception of the physical. However, this leads us back to the dilemma considered above. Therefore, we must see whether we can formulate an intelligible version of the knowledge argument without abandoning a theory-based conception of being physical.

An intelligible and not immediately false version of physicalism in the philosophy of mind can be tied to contemporary physics,[19] without endorsing the strong yet implausible claim that our contemporary physical knowledge provides the ultimate catalogue of physical entities. This "modest" formulation of physicalism relies on two assumptions. The first is that the current physics of "ordinary matter" is complete. The second is that the scientific study of the mind can be appropriately related to the contemporary scientific study of ordinary matter. Let us clarify these two claims.

The first assumption of this "modest formulation" of physicalism implies that a class of macroscopic phenomena can be completely described and explained in terms of the principles and properties of current physics.

---

[16] Levine (2001: 21).

[17] David Papineau illustrates how the idea of the completeness of physics is not a methodological or metaphysical principle based on *a priori* considerations. He argues that advancements in the understanding of neurophysiology due to biochemistry in the first half of twentieth century are central to establishing this principle. See Papineau (2002: 232–256).

[18] See Fodor (1974), Smart (1959) and Churchland (1986).

[19] A proposal of this type is forcefully presented in Smart (1978) and Smart (1989).

Jack Smart illustrates this by referring to the position of the physicist Gerald Feinberg:[20]

> [T]he theory of the electron, proton, neutron, neutrino and photon and their anti-particles, when they have such, is enough to explain the properties of ordinary matter (not what goes on inside neutron stars or inside black holes, or the behaviour of the transitory particles created only with big cyclotrons). Feinberg thus holds that the "Thales Problem" (of what the world of familiar objects is made of) has essentially been solved. (Smart 1989: 81)

Thus we should endorse Smart's claim that changes in theories concerning phenomena at the sub-atomic level, which are studied under special laboratory conditions, can be expected. Moreover, we can expect transformations in those theories that consider the universe as a whole. However, similar changes will not affect scientific descriptions and explanations of macroscopic phenomena involving ordinary matter. There will be no discoveries altering the fact that the hydrogen atom contains one proton and one electron, or that water is H_2O.

The second assumption of the "modest" formulation of physicalism establishes a programme for unifying the science of the mind with that of ordinary matter. Physicalists have notoriously disagreed on the proper formulation of this programme. Some articulate it in reductive terms; specifically, they assume that psychology will be reduced to neuroscience, given that, in general, each type of mental state is identical to a certain type of brain state.[21] Other physicalists or naturalists combine the thesis of the ontological primacy of the physical over the mental with some notion of the autonomy of psychology from neuroscience and physics.[22] Thus, instead of endorsing reductionism, they claim that each token of a type of mental state is identical to a certain physical token. Moreover, they argue that relations of supervenience or emergence which are weaker than identity exist between types of mental states and certain types of physical states. Insofar as we assume that Mary knows the relevant scientific disciplines involved in the physicalist unification programme, we need not decide on this issue.

The "modest" formulation of physicalism appears to be a plausible and intelligible general version of physicalism in the philosophy of mind. However, as it stands, it does not offer a precise enough characterisation of Mary's scientific knowledge. Clearly, Dennett's concerns cannot be addressed by assuming that Mary knows about the properties of conscious

[20] Feinberg (1966).

[21] Smart (1959) and Hill (1991).

[22] Fodor (1974) and Putnam (1967). Dretske (1995) offers a non-reductive naturalistic account of colour experience.

experiences investigated by unspecified psychological doctrines. Moreover, claiming that these doctrines are to be unified with the study of the macroscopic physical properties of the brain leaves huge gaps in our understanding of her knowledge. We need, then, to examine whether a more detailed characterisation of Mary's scientific knowledge can be provided within the general framework of modest physicalism.

4. A Scientific Categorisation of Colour Experiences

The knowledge argument is supposed to reveal the existence of *qualia*, understood as non-physical properties that typify conscious experiences. Therefore, two colour experiences are said to be of the same type if they share the same *quale*. If we wish to formulate the knowledge argument as a challenge to modest physicalism, this suggests how the problem of accounting for Mary's scientific knowledge might be addressed. The idea is to investigate whether we can ascribe to her a *complete knowledge* of the contemporary psychometric methods employed in describing the features that ground the categorisation of colour experiences.

In contemporary science, *spatial models* are used to describe how we categorise different types of stimuli. Here we are interested in colour spaces that describe our categorisation of colours.[23] I will refer to a *colour solid* as one such colour space. A colour solid can represent the ordering of colours that we discriminate by means of three dimensions: *hue*, *saturation* and *lightness* or *brightness* (see Figure 1).

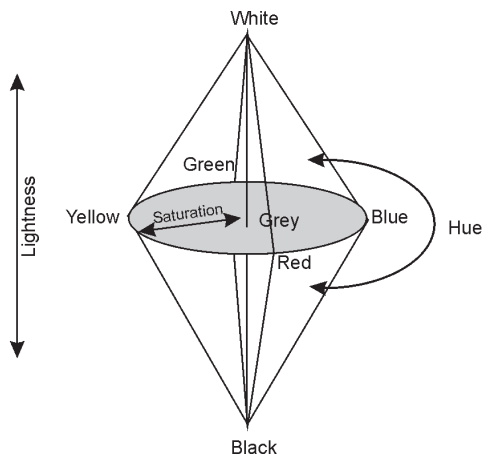

Figure 1 – *A colour solid*

[23] The central role of quality spaces in contemporary colour vision science is explored in detail in Clark (1993) and Clark (2000). See also Palmer (1999a).

Hue is the dimension we normally associate with the basic colours of surfaces. In the colour solid, hue is represented by the angular direction in the horizontal plane, from the central axis of the solid to the point position representing that colour. Saturation is the "vividness" of a colour. Chromatic colours of the same hue can differ in strength of hue, with less saturated colours closer to grey than more saturated ones. In the colour solid, the saturation of a certain colour is represented by the distance of the point representing that colour from the colour solid's central axis. The third dimension of colour is brightness, i.e. the relative lightness or darkness of a particular colour, from black to white.[24] In the colour solid, the brightness of a certain colour is represented by the height of the point representing that colour. The colour solid encodes information about the colours that subjects discriminate. Points in this space are taken to represent shades of colour, as these points are individuated by specific values in the axes representing hue, saturation and lightness. In addition, the relative distances of points representing shades of colour define the relative similarities between those shades. For example, orange would be situated between red and yellow, because subjects find it more similar to these two colours than, say, to blue.[25]

Let us now consider how colour spaces are determined. There are experimental psychophysical procedures for determining how individuals categorise colours. These consist in stimulating subjects' visual system and registering their discriminatory responses. The stimuli employed are lights, which are characterised physically in terms of the wavelength and intensity of the electromagnetic waves composing them. Discriminatory responses are observable behaviours; scientists usually examine verbal reports, but other kinds of behavioural clues can also be used. In particular, psychophysicists use a notion of indiscriminability that satisfies some statistically defined conditions, such that establishing whether two stimuli are indiscriminable requires the reiterated presentation of pairs of stimuli of the same type. Thus, if different individuals with normal vision fail

[24] Generally the term "brightness" is used to indicate a feature of colours seen through apertures or those of self-luminous objects like the sun or lamps. In contrast, "lightness" refers to a feature of colours of objects not seen through apertures or perceived as self-luminous.

[25] The colour solid also encodes information about relations of *composition* between colours. Certain hues can be analysed in terms of which hues are more basic. Orange, for instance, appears to contain both redness and yellowness. For this reason, a certain shade of orange will be represented by a location between red and yellow. In contrast, particular shades of red, green, blue and yellow do not appear to be composed of any colours. The colour solid also represents relations of opponency between colours. For example, it shows that there are no hues which appear reddish and greenish. In fact, there is no point in the colour space that might represent such hues.

to notice any difference between two stimuli in a statistically significant way, the two stimuli are said to be indiscriminable. These procedures determine the classes of indiscriminable stimuli.

The colour solid is obtained using mathematical procedures applied to classes of indiscriminable stimuli determined by considering subjects' judgments concerning the similarity of these stimuli. These methods must define the number of dimensions of the colour solid and the structure of relations of similarity between its points. Different methods have been investigated to this end.[26] A family of statistical procedures known as *multidimensional scaling* (MDS) has been successfully employed in psychophysics.[27] We can illustrate these techniques with an example. In colour science, multidimensional scaling can be applied to similarity matrices representing a relation of similarity among stimuli. Table 1 shows a matrix representing similarity ratings of monochromatic light stimuli described in terms of their wavelength.[28] For each pair of stimuli, a numerical value represents their degree of similarity. These values have been determined experimentally mainly by registering subjects' discriminatory judgements.

Wavelength	445	465	504	537	584	600	651	674
445	–	9	7	6	2	2	7	8
465		–	8	7	2	2	6	7
504			–	9	6	5	2	2
537				–	7	6	3	2
584					–	8	4	3
600						–	5	4
651							–	9
674								–

Table 1 – *A similarity matrix for observers based on* Ekman 1954.

[26] Some philosophers have investigated these procedures. Rudolf Carnap, for instance, in his attempt to provide a method for constituting all scientific concepts from a base of observational primitives, faced the problem of determining colour classes originating from classes of pairs of certain primitive particulars; Carnap (1967: 107–136, 178–182). Nelson Goodman subsequently demonstrated some limitations in Carnap's methods, and developed an alternative approach; see Chapter 5 of Goodman (1977). For a comparison of these two approaches, see Clark (1993: 101–112).

[27] The use of MDS for determining the colour space was advocated by Shepard (1962). An introduction to MDS is offered in Clark (1993: 210–221), while a more exhaustive and technical presentation can be found in Shifman et al. (1981).

[28] Monochromatic stimuli are those characterised by a single wavelength.

Using matrices of this type, a spatial model is determined by applying the algorithms of multidimensional scaling.[29] Figure 2 shows a map obtained by applying MDS to the colour similarity data in Table 1. The MDS procedure shows that the similarities between the stimuli concern only two dimensions: saturation and hue. This is because the initial matrix is not complete, and does not contain enough information to derive the dimension of brightness. In principle, however, if a complete matrix of similarity is available, the structure and dimensionality of a complete colour space are obtainable in the same way.

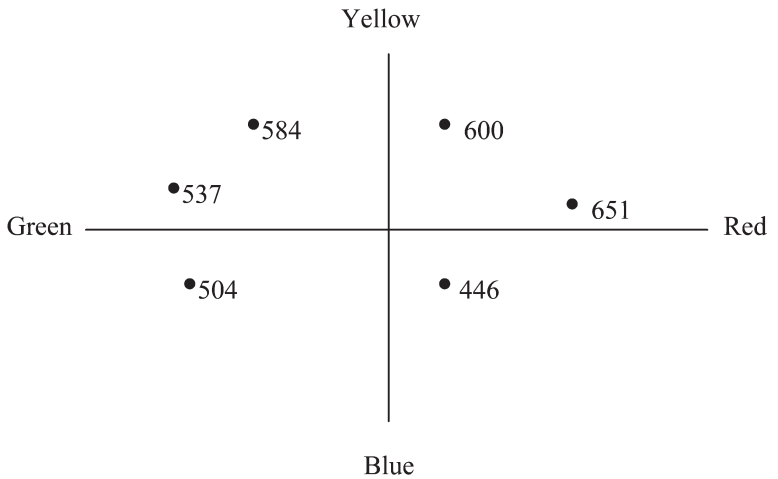

Figure 2 – *Multidimensional scaling map based on the colour similarity data in Table 1*. (Coren, Ward, and Enns 1999: 46)

To sum up, we have seen that, in principle, it is possible to use classes of discriminatory responses to light stimuli in order to determine the dimensions of variations involved in our classification of colours. In particular, we have seen that each colour can be individuated and described as a certain position in a colour space. However, we still need to specify how contemporary science can describe colour experiences by means of colour spaces.

[29] The procedure illustrated here is an instance of *metric* multidimensional scaling. In fact, the matrix under consideration presents the degree of similarity between the responses evoked by two colour stimuli in terms of a specific numerical value. However, it is also possible to generate a colour space by applying *non-metric* MDS procedures. In this case, the matrix does not contain values concerning the degree of similarity between stimuli. More specifically, it is possible to obtain a colour space by applying non-metric MDS to matrices representing judgements involving the triadic relation "*x* is more similar to *y* than *z*"; see Clark (1993: 220–221).

5. Accounting for Colour Experiences

Colour spaces do not seem to advance our understanding of how contemporary science aims to describe the features that ground the categorisation of colour experiences. Thus we are left with the problem of providing a substantive account of how Mary might describe such experiences.

Surely, any commitment to a theory of colour should be avoided. In fact, views on the relation between colours and colour experiences vary. Some consider colours to be properties of experiences or other mental items,[30] while others regard colours as properties whose instantiation in an object is identical to the power of that object to produce a colour experience of a certain type.[31] Finally, some consider colours to be objective properties of external objects.[32] The question is whether colour spaces and colour experiences can be correlated in a way that is relatively neutral with regard to these theories of colour.

Coloured objects appear to subjects in certain ways. In the case of a subject seeing a red patch, this patch appears to her to have certain features; in particular, it looks red to her. Now, some similarity is involved in seeing an object of a certain colour and an object that merely looks like it has that colour. For instance, colour illusions may be caused by changes in illumination, or by simultaneous contrast. An example of the first type of illusion is found when viewing a red patch under yellow light; in this case, the patch will appear visually to be blue, i.e. it seems to the subject to have the same colour as a blue patch. Likewise, in a standard demonstration of simultaneous contrast, a grey square on a red background looks greenish, while the same square seen against a green background looks reddish.[33] In both cases, it may be said that the square seems to have the same colour as a greenish or a reddish square, respectively.[34] Thus, even in cases of illusory perception, objects appear to have a certain colour.

Colour spaces describe the colour that objects seem to have. The previous section showed that colour spaces are determined by considering discriminatory responses to colour stimuli. A central assumption in psychophysics is that if certain individuals cannot discriminate between two

[30] Arguments against identifying colours with physical properties of the stimuli can be found in Hardin (1988).

[31] See McGinn (1991) and Peacocke (1984).

[32] See Byrne and Hilbert (2003) and Tye (2000).

[33] See Palmer (1999b).

[34] Assuming that there is such a similarity does not involve taking a position on the nature of colour. I am not speaking of whether the colour of an object is the colour that it *appears to have visually* in certain conditions. In addition, one need not assume that colours are properties of colour experiences.

stimuli, then these stimuli will seem to them to have the same colour.[35] Moreover, using matrices that arrange colour stimuli in accordance with the discriminatory responses they elicit, it is possible to determine the dimensions involved. In the case of a colour solid, for example, these dimensions are saturation, hue, and brightness. These dimensions are derived from discriminatory responses assumed to reflect relations of similarity and difference in the ways that colours appear to subjects. Therefore, they can be interpreted as dimensions of variation in colour. In particular, specific values for these dimensions specify a location in the colour space, which can be regarded as a description of the colour an object visually appears to have.[36]

Besides the colours objects appear to have, we can also assume that there are *qualitative properties* of colour experiences.[37] Qualitative properties individuate colour experiences. Thus if two colour experiences differ, they have different qualitative properties. It seems that we can establish an important relationship between qualitative properties and colours.

Some authors have argued that colour experiences can be categorised in terms of their typical causes.[38] A colour experience is of a certain type or, in our terminology, it has a certain qualitative property, if a certain paradigmatic object would produce it under certain conditions. For example, we can say that a *red-type* colour experience is the type produced by a certain paradigmatically red thing in certain suitable circumstances.

It has also been suggested that when something appears to the subject to be a certain colour, either in a perceptual case or in an illusory one, then this subject has a certain type of colour experience. Thus a subject can have a colour experience that is not produced by the paradigmatic object involved in characterising that type of colour experience. In such cases, the subject is said to be having this type of colour experience because the stimulus looks like the paradigmatic object. For example, if the subject

[35] More precisely, it is assumed that if two stimuli are *globally indiscriminable* (by certain individuals in certain conditions), then they appear to be the same colour (to those individuals in those conditions). A certain stimulus x is globally indiscriminable from stimulus y, if and only if for any stimulus z, x is indiscriminable from z if and only if y is as well. The reasons for adopting the notion of global indiscriminability in order to characterise the relation of appearing to be the same colour are discussed by Clark (1993: 56–62).

[36] Introducing the notion of a colour that an object appears to have visually does not involve any philosophical stance regarding this appearance. Whatever this appearance may turn out to be, here I am only showing that there is a way to provide a description of it by analysing subjects' discriminatory responses.

[37] This terminology is provided by Strawson (1989). See also Sellars (1963: 93–94, 192-93) and Clark (2000: 6).

[38] A typology of this kind is suggested by Peacocke (1984: 349–350), and further elaborated by Millar (1991: 25–31).

sees a grey patch on a yellow background, we would say that he is having the type of experience that a reddish patch would produce in certain circumstances.

The typology I have just described suggests a correspondence between the colour that an object appears to have and the qualitative property that specifies a type of colour experience. In this account, if a stimulus looks to the subject to be the same colour as a certain paradigmatic stimulus, then the subject has the type of colour experience that would be produced by the paradigmatic stimulus. Thus an individual *S* has a colour experience with a certain qualitative property when something appears to *S* to have the same colour as the paradigmatic stimulus.

The descriptions provided by a colour space can be used to categorise colour experiences; indeed, the points in the colour space represent the colours that objects appear to have. Thus, for every colour described by the colour space, one can associate a corresponding description of a certain type of colour experience. For instance, let us assume that a certain shade is represented in a colour space by the position *XYZ*.[39] The description *XYZ* specifies values of hue, saturation and brightness in terms of a system of relations of similarity with other colours. The relative type of colour experience will, therefore, be described as the type of experience someone has when something looks *XYZ* to him.

To sum up, it seems we can provide a clear enough account of how contemporary science categorises colour experiences and describes the properties that ground this categorisation. However, we still need to establish whether these descriptions can help in formulating an intelligible version of the knowledge argument that can be used against a plausible, relevant version of modest physicalism.

6. Mary's Scientific Knowledge

In order to characterise Mary's scientific knowledge, some remaining issues must be considered. First of all, we need to determine how the categorisation of colour experiences discussed above might fit into the modest physicalist programme I recommended in section three. This requires showing how these descriptions can be unified with the study of the macroscopic physical features of the brain. Contemporary science offers some hints.

[39] Each term referring to a shade of colour can be defined by a description which refers only to the relations of similarity (and of opponency and composition) represented by the colour space. This can be achieved using a logical technique that involves "Ramsey sentences". For more details, see Clark (2000: 256–257).

Scientists aim to explain the structure of colour quality spaces by means of neuro-physiological mechanisms that generate the relations between points in a quality space.[40] For example, the *opponent processors* theory explains the structure of a colour solid, and thereby the location of any colour shade within it, in terms of the activity of certain neural mechanisms.[41] Groups of neurons, by virtue of their excitatory and inhibitory connections, manifest regular response patterns to the outputs of three different types of photoreceptors in the retina. These opponent processors have a positive response to stimuli in a certain part of the spectrum, and a negative response to those in other parts of it. Based on certain empirical evidence, it is assumed that there are two chromatic opponent processors (blue-yellow and red-green) and one achromatic one (white-black). These three opponent processors generate the three axes of the colour solid. In fact, the different activation of each of these processors, resulting from a determinate stimulation of the photoreceptors in the retina, defines the position of a certain shade with respect of each of the solid's axes. A very bright orange, for example, will result from combined activation of the neural processes correlated with the experience of red, yellow and white and inhibition of the correlates of green, blue and black.

So far we do not have an account of Mary's scientific knowledge which is specific enough to offer an intelligible version of the knowledge argument. In fact, we have some idea of how the categorisation of colour experiences by means of certain colour spaces can be a descriptive requirement for a modest physicalist programme in the philosophy of mind. Yet although we can grasp the programmatic lines of this position, we may be concerned that even our knowledge of the brain's macroscopic features and processes is far from complete. However, it has been seen that we might have a satisfactory grasp of a scientific way of describing the features that ground a categorisation of colour experiences. Such descriptions offer the basic data that neuroscience is assumed to be able to explain. Thus we could use a version of the knowledge argument to test a substantive formulation of physicalism, by focusing on the capacity of this descriptive apparatus to account for properties that ground the categorisation of colour experiences.

Approaching this task requires some further clarification, as the determination of colour solids is far from complete at present. One reason

[40] In ordinary scientific practice, the phenomena described by psychophysics are often used to postulate relative neural mechanisms. On the general logical structure of these inferences, see Teller (1984).

[41] Hurvich (1981) offers a comprehensive and detailed presentation of this theory by one of its most important advocates. See also De Valois and De Valois (1975: 100–110). Simplified accounts of the theory can be found in Hardin (1988) and Clark (1993).

is that the set of discriminatory judgements required to achieve this aim is so large that its study is technically impossible.[42] Secondly, it might not be the case that a three-dimensional space, such as a colour solid, would provide an exhaustive description of colour experiences.[43] In fact, a three-dimensional colour solid is obtained by examining colour stimuli seen under specific laboratory conditions. However, as Clark points out, "glossy surfaces, reflections, translucency, transparency, shadows, and mists all require dimensions of variations in appearance beyond the three sufficient for coloured surfaces or lights presented in the lab" (Clark 2000: 7). Nonetheless, the general nature of the methods which might be used to offer a complete description of the colours that objects appear to have is clear enough. The scope of this approach is to determine, using a set of discriminatory judgements, the sensory dimensions by which subjects can discriminate colours. Once these dimensions are determined, it is possible to derive a qualitative space.

We can now assume that Mary knows how colours look to subjects in terms of their positions in a completed qualitative space. We can, therefore, ascribe to her a certain type of intelligible scientific knowledge. Mary can have complete scientific knowledge without having had colour experiences. The main notions involved in the psychophysical categorisation of colour experiences are provided by a colour space obtained from certain statistical procedures using discriminatory judgements concerning certain stimuli. It seems that none of these notions requires actually undergoing colour experiences in order to be completely understood. A colour space is a geometrical representation of the dimensions along which subjects discriminate colours. We can concede that Mary has an n-dimensional model of the different colours. Moreover, she knows the typology of colour experiences derivable from such a colour space. This model gives her information about the dimensions in which light stimuli are categorised by the visual system. Understanding that the visual system enables certain kinds of discrimination along these dimensions does not seem to require actually having colour experiences. Mary will know that these dimensions result from the application of certain statistical methods to sets of discriminatory responses, or judgements of similarity, elicited by certain physical stimuli. In particular, similar procedures can tell scientists about the dimensions along which a certain species' sensory system cat-

[42] For example, if we consider just 20 stimuli, 190 rankings of similarities among pairs are needed to fill out the data matrix for determining a relative quality space. Each such ranking may require many trials. It is assumed that human subjects can discriminate ten million colours; therefore, the determination of only one percent of this space would require five billion similarity rankings. See Clark (1993: 118).

[43] See Clark (2000: 39).

egorises certain stimuli. In particular, these procedures can be applied to species with sensory systems that human beings do not share, such as, for instance, those involved in echolocation.

Understanding the notion of a discriminatory response in psychophysics does not require having colour experiences. Mary understands that these responses are observable behaviours, such as a subject's verbal reports on the similarities and differences between certain physical stimuli. In particular, she knows that these responses have to satisfy certain statistical conditions in order to count as reliable evidence of subjects' discriminatory capacities. Nevertheless, none of these requirements implies that she cannot exhaustively observe these responses on a black-and-white screen.

Given the statistical nature of the psychophysical investigation of colour categorisation, stimuli must be repeatable types. Mary, in line with contemporary science, can regard these types as defined by physical properties described in the electromagnetic theory of light. In particular, stimuli are characterised in terms of the intensity and wavelength composition of the light imaged on the retina. Mary can have an understanding of these features from her reading of physics books. In addition, she can detect and measure them with instruments which we already possess, whose use does not require having colour experiences.

Finally, we can formulate a version of the knowledge argument which represents an intelligible objection to a substantial formulation of a physicalist programme within the boundaries of contemporary science. The first premise of this formulation of the knowledge argument is that, before her release, Mary can refer to the properties that ground the categorisation of colour experiences using the descriptions offered by completed colour spaces. The second premise states that Mary, by seeing a coloured object after her release, acquires *new knowledge* about colour experiences. This knowledge concerns the occurrence of *qualia*, i.e. features of colour experiences which, from the perspective acquired by seeing colours, ground a categorisation of colour experiences. Thus a generalisation of the descriptive apparatus of the contemporary science of colour vision cannot accommodate *qualia*, and the modest programme that I have delineated as a plausible and relevant version of physicalism is, therefore, untenable.

Establishing whether this version of the knowledge argument is sound must be left for another occasion. I will limit myself here to showing that this reasoning is immune to Dennett's "blue banana trick". As seen in the second section of this paper, Dennett suggested that Mary's scientific knowledge might enable her to know what it is like to see colours prior to her release. In particular, she would be able to recognise on her release that a blue banana has the wrong colour. However, this is not plausible in

the version of the knowledge argument that I have advanced in this paper. Before her release, Mary knows how blue and yellow things look in terms of the positions of yellow and blue in the system of relations of similarity embedded in the complete colour space. On seeing a blue banana, however, Mary can have only very limited relational information about this object's colour. The only relations of similarity and difference she might actually discriminate are those between the banana's appearance and the background. Such evidence is insufficient for her to know which colour she is seeing, and hence the type of colour experience she is having.

7. Conclusion

I have endorsed Daniel Dennett's criticism of Frank Jackson's version of the knowledge argument, according to which Jackson makes use of an unintelligible premise concerning Mary's complete scientific knowledge of colour experiences. However, I have also argued that we should not abandon the strategy implicit in the knowledge argument, offering a substantial and understandable formulation of Mary's relevant knowledge of colour experiences. This knowledge fits with a plausible and intelligible version of physicalism in the philosophy of mind. Whether the resulting version of the knowledge argument is sound remains to be seen.

References

Alter, T. 1999. "The Knowledge Argument", available at: http://host.uniroma3.it/progetti/kant/field/ka.html

Byrne, A. and Hilbert, D. 2003. "Color Realism and Color Science", *Behavioral and Brain Sciences* 26: 3–64.

Carnap, R. 1967. *The Logical Structure of the World* (Berkeley: University of California Press).

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

Churchland, P.S. 1986. *Neurophilosophy: Toward a Unified Science of the Mind-Brain* (Cambridge, Mass.: MIT Press).

Clark, A. 1993. *Sensory Qualities* (Oxford: Clarendon Press).

—— 2000. *A Theory of Sentience* (Oxford: Oxford University Press).

Coren, S., Ward, L.M. and Enns, J.T. 1999. *Sensation and Perception* (5th edition, London: Harcourt Brace College Publishers).

Crane, T. and Mellor, D.H. 1990. "There is No Question of Physicalism", *Mind* 99: 185–206.

Davidson, D. 1970. "Mental Events", in: Foster, L. and Swanson, J.W. (eds.), *Experience and Theory* (Amherst, Mass.: University of Massachusetts Press) 79–91. Reprinted in Davidson, D. 2001. *Essays on Actions and Events* (2nd edition, Oxford: Oxford University Press) 207–225.

De Valois, R.L. and De Valois, K.K. 1975. "Neural Coding of Color", in: Carterette, E.C. and Friedman, M.P. (eds.), *Handbook of Perception, Vol. 5: Seeing* (New York: Academic Press) 117–66. Reprinted in Byrne, A. and Hilbert, D.R. 1997. (eds.), *Readings on Color. Volume Two: The Science of Color* (Cambridge, Mass. and London: MIT Press) 93–140.

Dennett, D. 1988. "Quining Qualia", in: Marcel, A. and Bisiach, E. (eds.), *Consciousness in Contemporary Science* (Oxford: Oxford University Press) 43–77. Reprinted in Block, N., Flanagan, O. and Güzeldere, G. 1997 (eds.), *The Nature of Consciousness* (Cambridge, Mass.: MIT Press) 619–642.

—— 1991. *Consciousness Explained* (London: Little & Brown; reprinted, London: Penguin, 1993).

Dretske, F. 1995. *Naturalizing the Mind* (Cambridge, Mass.: MIT Press).

Ekman, G. 1954. "Dimensions of Color Vision", *Journal of Psychology* 38: 467–474.

Feinberg, G. 1966. "Physics and the Thales Problem", *Journal of Philosophy* 63: 5–17.

Fodor, J. 1974. "Special Sciences (or The Disunity of Science as a Working Hypothesis)", *Synthese* 28: 97–115. Reprinted in Block, N. 1980 (ed.), *Readings in Philosophy of Psychology. Volume One* (Cambridge, Mass.: Harvard University Press) 120–133.

Goodman, N. 1977. *The Structure of Appearance* (Dordrecht and Boston: Reidel).

Hardin, C.L. 1988. *Color for Philosophers: Unweaving the Rainbow* (Indianapolis and Cambridge: Hackett Publishing Company).

Hempel, C.G. 1980. "Comments on Goodman's Ways of Worldmaking", *Synthese* 45: 139–99.

Hill, C. 1991. *Sensations: A Defense of Type Materialism* (Cambridge: Cambridge University Press).

Hurvich, L. 1981. *Color Vision* (Sunderland, Mass.: Sinauer Associates Inc.).

Jackson, F. 1982. "Epiphenomenal Qualia", *Philosophical Quarterly* 32: 127–36. Reprinted in Lycan, W. 1990 (ed.), *Mind and Cognition* (Oxford: Blackwell) 469–77.

—— 1986. "What Mary Didn't Know", *Journal of Philosophy* 83: 291–5. Reprinted in Block, N., Flanagan, O. and Güzeldere, G. 1997 (eds.), *The Nature of Consciousness* (Cambridge, Mass.: MIT Press) 567–570.

—— 1998. “Postscript on Qualia”, in: Jackson, F. *Mind, Method and Conditionals: Selected Essays* (London: Routledge) 76–79.

—— 2004. “Mind and Illusion”, in: Ludlow, P., Nagasawa, J. and Stoljar, D. (eds.), *There’s Something about Mary* (Cambridge, Mass.: MIT Press) 421–442.

Levine, J. 2001. *Purple Haze: The Puzzle of Consciousness* (Oxford: Oxford University Press).

McGinn, C. 1991. *The Problem of Consciousness* (Oxford: Blackwell).

Mellor, D.H. 1993. “Nothing Like Experience”, *Proceedings of the Aristotelian Society* 93: 1–16.

Millar, A. 1991. *Reasons and Experience* (Oxford: Clarendon Press).

Montero, B. 1999. “The Body Problem”, *Nous* 33: 183–200.

Nida-Rümelin, M. 2002. “Qualia: The Knowledge Argument”, in Zalta, E.N. (ed.), *The Stanford Encyclopaedia of Philosophy*, http://plato.stanford.edu/entries/qualia-knowledge

Palmer, S.E. 1999a. “Color, Consciousness and the Isomorphism Constraint”, *Behavioral and Brain Sciences* 22: 923–989.

—— 1999b. *Vision Science* (Cambridge, Mass.: MIT Press).

Papineau, D. 2002. *Thinking about Consciousness* (Oxford: Clarendon Press).

Peacocke, C. 1984. “Colour Concepts and Colour Experience”, *Synthese* 58: 365–82.

Putnam, H. 1967. “Psychological Predicates”, in: Capitan, W.H. and Merrill, D.D. (eds.), *Art, Mind and Religion* (Pittsburgh: University of Pittsburgh Press). Reprinted as “The Nature of Mental States” in Putnam, H. 1975. *Mind, Language, and Reality. Philosophical Papers* Vol. 2 (Cambridge: Cambridge University Press) 429–440.

Robinson, H. 1993. “Dennett on the Knowledge Argument”, *Analysis* 53: 174–77.

Sellars, W. 1963. *Science, Perception, and Reality* (London: Routledge & Kegan Paul).

Shepard, R.N. 1962. “The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function: Part I”, *Psychometrika* 27(3): 125–40.

Shifman, S.S., Reynolds, M.L. and Young, F.W. 1981. *Introduction to Multidimensional Scaling* (New York: Academic Press).

Smart, J.J.C. 1959. “Sensations and Brain Processes”, *The Philosophical Review* 68(2): 141–156. Reprinted (revised) in Borst, C.V. 1970 (ed.), *The Mind/Brain Identity Theory* (London: Macmillan) 52–66.

—— 1978. “The Content of Physicalism”, *The Philosophical Quarterly* 28: 239–41.

—— 1989. *Our Place in the Universe* (Oxford: Oxford University Press).

Strawson, G. 1989. "Red and 'Red'", *Synthese* 78: 193–232.

Teller, D.Y. 1984. "Linking Propositions", *Vision Research* 24(10): 1233–1246.

Tye, M. 2000. *Consciousness, Color, and Content* (Cambridge, Mass. and London: MIT Press).