Why Do Colours Look the Way They Do?

NICHOLAS UNWIN

Abstract
A major part of the mind–body problem is to explain why a given set of physical processes should give rise to perceptual qualities of one sort rather than another. Colour hues are the usual example considered here, and there is a lively debate as to whether the results of colour vision science can provide convincing explanations of why colours actually look the way they do. The internal phenomenological structure of colours is considered here in some detail, and a comparison is drawn with sounds and their synthesis. This paper examines the type of explanation that is needed, and it is concluded that it does not have to be reductive to be effective. What needs to be explained more than anything is why inverted *hue* scenarios are more intuitive than other sensory inversions: and the issue of physicalism versus dualism is argued to be of only marginal relevance.

1. Introduction

Colour perceptions have long been thought to present a major stumbling block to understanding the nature of the mental and how it relates to the physical. Seeing red and seeing green, for example, have often been supposed to be indistinguishable from a purely functional point of view, which means that it is impossible to rule out inverted perceptual qualities (or '*qualia*'), for example, the possibility that you might see red where I see green, and so on. Joseph Levine argues for an 'explanatory gap' between the mental and the physical: all the facts about light, the surfaces of physical objects, and the workings of the eye and the brain will not explain why colours actually look the way they do; why green-stimuli give rise to green *qualia* as opposed to red ones, for example.[1]

However, C.L. Hardin, Austen Clark and others have argued that the situation is not so bleak.[2] Colour vision science suggests a

[1] Joseph Levine, 'Materialism and Qualia: The Explanatory Gap' in *Pacific Philosophical Quarterly* **64** (1983), 354–61.

[2] C.L. Hardin, 'Qualia and Materialism: Closing the Explanatory Gap', *Philosophy and Phenomenological Research* **48** (1987), 281–98; *Color for*

doi:10.1017/S0031819111000234

number of important asymmetries between different colour hues, asymmetries which are both genuinely phenomenological – that is to say, which concern actual appearances – and which can also be given physiological groundings. The gap between the mental and the physical is therefore far narrower than the pessimists maintain.

These issues are much debated, and in highly intricate detail. However, the detail often obscures some crucial, broader questions, and I shall argue that the type of explanation that we are dealing with here has been largely misunderstood by all parties. Specifically, reductionism is neither warranted by the scientific evidence, nor required in order to give intelligible explanations of the kind we should be interested in. What we get from colour vision science, and all we really need, are illuminating *connections*, not reductions, between mental and physical phenomena, and – more importantly – between the mental phenomena themselves; and this suggests, among other things, that the physicalism/dualism debate is largely irrelevant here, contrary to most current received opinion.

2. Some Results from Colour Vision Science

What sort of hue asymmetries are there? There are four reasonably well known examples. Firstly, it is now generally accepted that there is a real phenomenological distinction between *unique* and *binary* hues. The unique hues (red, yellow, green, blue) look essentially unmixed, whereas the binary hues (orange, purple, turquoise, chartreuse) look essentially like mixtures of unique hues (yellow–red, red–blue, blue–green, green–yellow, respectively).[3] These purely

Philosophers: Unweaving the Rainbow (expanded edn) (Indianapolis, IN: Hackett, 1988); 'Reinverting the Spectrum', in Alex Byrne and David R. Hilbert (eds), *Readings on Color, Vol. 1: The Philosophy of Color,* (Cambridge MA: MIT Press, (1997), 289–301; Austen Clark, *Sensory Qualities,* (Oxford: Clarendon Press, 1993); 'I am Joe's explanatory gap', at http://selfpace.uconn.edu/paper/PGAP.HTM (1994).

[3] The notion of 'mixture' used here is purely phenomenal, and should not be confused with what happens when differently coloured lights or pigments are physically combined: such combinations are often surprising, and precisely because they do not correspond to purely phenomenal mixtures. It is likewise important not to confuse 'unique hues' with 'primary colours' (either additive or subtractive). These distinctions are unobvious, and many people – notably Brentano – claim that green actually looks like a mixture of yellow and blue; and I once taught a class where everyone insisted

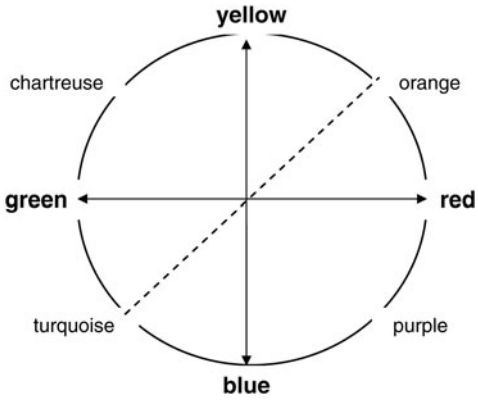
phenomenological facts have a well understood physiological grounding in Ewald Hering's opponent-process theory. That is to say, although there are only three kinds of colour photoreceptor in the eye – often described, misleadingly, as the red, green and blue cones–-differences in stimulation level group themselves into two retino-cortical channels: the red – green channel and the yellow–blue channel (plus the achromatic white–black channel). Thus unique hues correspond to the activation of just one channel, whereas binary hues correspond to the activation of two.

Secondly, and relatedly, there are some combinations of unique hues that are perceptually impossible, namely red–green and yellow–blue.[4] This phenomenological fact has its correlate in the fact that the red–green and yellow–blue channels are each (what in physiology are called) 'opponent processes'. When I see red (or a reddish hue), the red–green channel is excited; when I see green (or a greenish hue), it is inhibited (likewise with yellow and blue). A given channel cannot be simultaneously excited and inhibited, any more than an energy level can simultaneously increase and decrease, and this explains the phenomenal opponences. The relevant phenomenological facts may be summarized in the following diagram:

---

that they could see a unique purple which looked neither reddish nor bluish. Introspective reports are notoriously unreliable; however, more precise psychophysical experiments have largely stabilized the discussion. See, for example, Leo M. Hurvich, 'Chromatic and Achromatic Response Functions', in Alex Byrne & David R. Hilbert (eds), *Readings on Color, Vol. 2: The Science of Color*. Cambridge MA: MIT Press (1997).

[4] Except in extraordinary, laboratory-induced circumstances. On this, see H. Crane and T.P. Piantanida, 'On Seeing Reddish Green and Yellowish Blue', *Science* **221** (1983), 1078–80; V.A. Billock, G.A. Gleason and B.H. Tsou, 'Perception of forbidden colors in retinally stabilized equiluminant images: an indication of softwired cortical color opponency?', *Journal of the Optical Society of America A* **18** (2001), 2398–403; and Juan Suarez and Martine Nida-Rümelin, 'Reddish Green: A Challenge for Modal Claims about Phenomenal Structure', *Philosophy and Phenomenological Research* **78** (2009), 346–91. The circumstances involve the use of filling-in mechanisms, which occur within the visual cortex itself, and therefore, arguably, do not undermine the opponent process theory, which concerns only retino-cortical channels. That unusual phenomenology should result from unusual stimuli (and brain processes), if anything, rather tends to support the view that hue phenomenology is explicable in physiological terms; so these phenomena, although remarkable, are not relevantly embarrassing.



Thirdly, and much more controversially, red and yellow look intrinsically 'warm' ('positive' or 'advancing') whereas green and blue look intrinsically 'cool' ('negative' or 'receding'). This may be purely cultural, or due to familiar physical associations, but perhaps not. It might also be that there are actual physiological connections between opponent channel excitation and those neurons implicated in the sensation of warmth, and likewise between opponent channel inhibition and those neurons implicated in the sensation of coolness.[5]

Fourthly, yellow is a much lighter colour than red, and this is easily explained by the fact that yellow appears in the middle of the visible spectrum, where the light–dark achromatic response curve peaks. Further asymmetries include the greater number of perceptible differences between red and blue than between yellow and green.[6] This might be an immediate consequence of the lower chromatic

[5] Peter K. Kaiser surveys some results here in 'Physiological Response to Color: A Critical Review', *Color Research and Application* **9** (1984), 29–36. There is, alas, not much evidence of the sort of neural links that we require, though it is sufficient, for philosophical purposes, to ask what would follow if there were. A further complication is that the warm/cool distinction is also connected to achromatic light/dark: on this, see B. Berlin and P. Kay, *Basic Color Terms: Their Universality and Evolution*, (Berkeley CA: University of California Press, 1969).

[6] There are many results that support this, some of them directly psychophysical, some of them anthropological: the variations in colour vocabulary between different cultures has, since Berlin & Kay (1969), proved to be a very fertile research tool in this area: see, for example, C.L. Hardin and Luisa Maffi (eds), *Color Categories in Thought and Language*, (Cambridge: Cambridge University Press, 1997) for a useful survey.

content of yellow (and, to a lesser extent, green) compared to red and blue, and therefore not a phenomenon independent of the chroma point; though nobody is quite sure.

The upshot, it would seem, is that we have enough asymmetries to rule out any possible systematic inversion of the colour spectrum. If an alternative permutation is to be undetectable, then unique hues must map onto unique hues and binary hues onto binary hues, as is shown by the first point (that there should be such a distinction in the first place is also thus guaranteed, with the second point giving further details). Likewise warm hues must map onto warm hues and cool hues onto cool hues if we are to accommodate the third point, and this rules out the standard red–green inversion scenario. The only remaining possibility is what we may call 'diagonal inversion', i.e. a reflection in the dotted diagonal axis of the colour circle depicted above, where red is exchanged with yellow, and green with blue; and this is apparently excluded by the fourth point. It would seem, then, that undetectable hue-inversions can be completely ruled out.[7]

3. Asymmetry and Explanation

But how much does this actually *explain*? Do we really now understand why green looks like green as opposed to red? Or as opposed to blue? Levine argues against Hardin that we have not managed to traverse the explanatory gap at all: asymmetry does not *ipso facto* yield explanation.[8] I shall argue that Levine is right on the general point, but that the above connections yield more than bare asymmetries: at least, that is so for the first three points, though not the fourth. There are further distinctions here that need a more careful analysis.

To take the last point first, even if we take ourselves to have explained why red looks like red as opposed to yellow (since the latter is lighter than the former), this surely does nothing to explain why green looks like green as opposed to blue, since there are only very minor differences in chromatic content between green and blue. Secondly, and more generally, chroma (or saturation) and hue are

[7] For more on hue asymmetries, see Stephen E. Palmer *et al.*, 'Color, Consciousness and the Isomorphism Constraint' (plus commentaries), *Behavioral and Brain Sciences* **22** (1999), 923–89.

[8] Joseph Levine, 'Cool Red: A Reply to Hardin', *Philosophical Psychology* **4** (1991), 27–40.

generally understood to be quite independent dimensions of colour, and this independence is directly, visibly evident to us. It is therefore hard to see how facts about the former could ever really *explain* anything much about the latter. It follows that it is not even clear that the difference between red and yellow has been adequately explained, for why should not someone be able to perceive a 'supersaturated yellow', i.e. a colour that relates to yellow as red relates to pink? We cannot readily imagine such a thing; but it is nevertheless not an especially odd suggestion, for if the yellow hue is no longer at the peak of the light–dark curve, there is no longer any reason why it should be chromatically weak (whereas if the curve peaks instead at the red/pink part of the spectrum, as is required for diagonal-inverts, then 'supersaturated pink' – what we call 'red' – would become unperceivable). Likewise, the greater similarity between red and blue than between yellow and green might well automatically reverse as a result of these changes in chroma levels.[9]

Levine goes further still and attacks the third point, arguing that we could conceivably perceive a cool red or a warm green. If this is right, then Hardin's thesis is even more seriously damaged. It would seem that the asymmetries are nowhere near robust enough to sustain halfway decent explanations. Indeed, David Chalmers claims that such possibilities ensure that that there need be no useful hue asymmetries at all.[10]

Yet there are important differences, and it is here that I disagree with all parties. It is implausible that cool red should be treated as on a par with supersaturated yellow, for the warm/cool dimension, unlike saturation levels, really does seem to relate specifically to hue. We might indeed reasonably protest, with Hardin, that 'cool red' is a contradiction in terms, and that the feature which we naturally call 'warmth' is visibly an essential constituent of redness. Of course, it is not all of redness (otherwise yellow could not also be

[9] Since supersaturated yellow is virtually unimaginable to us, it can hardly resemble green as much as ordinary yellow does! Of course, this fact need not automatically carry over for diagonally-inverted percipients; but the point is that arguments about numbers of perceptible differences are clearly far less conclusive here than they are often thought to be. The possibility of a supersaturated yellow is discussed briefly by Hardin (1988), 140. 'Diagonal inversion' is examined more thoroughly in a PowerPoint presentation on my website, where digitally inverted photographs are used to illustrate the phenomenon. See http://www.lancs.ac.uk/fass/doc_library/ippp/unwin_ppt_why_do_colours_07.pdf

[10] David Chalmers, *The Conscious Mind: In Search of a Fundamental Theory*, (Oxford: Oxford University Press, 1996), 99–100.

warm), so the possibility remains that the residue, whatever it is, could be combined with coolness to produce a new hue. Yet this new hue would be seriously alien in a way in which supersaturated yellow surely is not: to obtain the latter, we merely have to stretch things a bit, but we have no idea what cool red could look like. Moreover, there is surely something intuitively satisfying about the putative connection between chromatic warm/cool with ordinary tactile warm/cool. There is a directly perceivable resemblance here. So, if (a very big 'if') a plausible neural route can be shown to connect the relevant visual and tactile parts of the brain, we have something that it is intuitively explanatory. We are surely right to feel that we have now learnt something important about why warm hues look warm and cool hues look cool. Inter-sensory connections go beyond mere asymmetries, and really do seem to yield genuine explanations. Furthermore, should it turn out that there are no such links after all, the conclusion should surely be, not that we have misidentified the shape of good explanations here, but rather that we just do not have any good explanations here.

Other sorts of connection are also explanatorily fertile. For example, the connection between *unique* and *binary* (hues) and (the activation of) *one* and *two* (opponent processing mechanisms) is very direct. It may also be that the *advancing/receding* distinction corresponds to how light of different wavelengths is focused within the eye itself; if so, this also provides an intuitive, demystifying relationship, and therefore at least the beginnings of an explanation.[11] The phenomenally penetrating character of red (as opposed to green, for example) likewise connects very naturally with the greater atmospheric penetrating power of red light.

How could we take this further? Inter-modal links are especially suggestive, and most people I have asked agree that yellows and greens have a quality that may be described as 'sharp', 'fresh' and 'citrusy', whereas reds, blues and purples do not. As with warmth and coolness, we need to guard against the possibility that we are merely dealing with familiar physical associations, but suppose that we can rule that out.[12] Suppose, further, that we can find convincing neural connections between the visual and gustatory centres of the brain. If so, then, I submit, we would once again have gained

[11] See Hardin (1988), 129.

[12] Other difficulties also include the sheer suggestibility of naïve experimental subjects. But there is no reason why more sophisticated experiments should not be made, perhaps with properly trained psychophysicists as subjects.

something genuinely explanatory. Unlike the chroma point, we really have now started to explain why red looks like red as opposed to yellow, and why green looks like green as opposed to blue; and the explanatory gap has therefore been very substantially reduced even further.

Why do inter-modal connections yield explanations whereas mere asymmetries do not? It is partly because such connections posit interesting facts about how the brain works, but this cannot be the whole story. Our original problem is that, even if we knew everything about the brain, we would still not understand the nature of colour *qualia*. No, the really significant development has been at the purely phenomenal level. In coming to notice inter-modal links (as well as the distinction between mixed and unmixed), we notice more about colour hues themselves. We can come to explain things about them, because we are getting clearer about just what it is about them that needs to be explained. One central obstacle to explaining the nature of colour appearances is that they appear to be ineffable, that is to say, indescribable and unanalysable: what Hume calls 'simple impressions'. What colour vision science might well do (and, to some extent, has already done) is to reveal underlying phenomenal structure. To begin with, we have the familiar colour dimensions of hue, saturation and brightness, each of which can be varied independently of the others. Hume's missing shade of blue was supposed to be in the middle of a sequence of blues all of the same hue but of increasing lightness (a combination of increasing brightness and decreasing saturation). Since lightness forms a straightforwardly linear scale, it is no particular mystery that we can successfully imagine this shade without ever having seen it. But what has been shown here is that a blue perception is not really a 'simple impression', even if the image is spatiotemporally homogeneous (I strongly suspect that there are no genuinely *simple* impressions at all, in any sensory mode). The hue/saturation/brightness dimensions already give it structure; and considerations of warmth and coolness, and so on, might uncover further structure within the hue dimension itself.

4. Sounds and Parameters

We can generalize this, and a useful comparison is with sounds. A musical synthesizer delivers a variety of sounds in a systematic way. Each sound can be edited by adjusting volume and pitch (as with an ordinary piano), but also a great variety of other 'parameters'

such as auditory brightness (i.e. how many harmonics are included),[13] attack and decay levels and many other things. A sustained single note is usually understood to have three dimensions, volume, pitch and timbre, just as a uniform colour has a given (visual) brightness, saturation and hue. Moreover, although volume and pitch, like brightness and saturation, form a straightforward linear scale, timbre (often called 'tone colour'), like hue, is much less easily characterized. We can all recognize the difference between the sound of a trumpet and the sound of a clarinet (at the same pitch and volume), but may find it difficult to put it into words. Now, the parametrization yielded by a synthesizer provides a common structure within which timbres can be located, so that each timbre can identified by the particular numerical values assigned to each parameter. Furthermore, as the name suggests, synthesizers can be used to create novel, hybrid sounds by choosing unusual values: for example, a sound with the attack of a piano but the sustain of a violin, or things more exotic still.

Now, suppose we consider a combination of values that we have never heard before: can we tell in advance what it will sound like? Often yes. When we hear the new sound, we may be entertained, but not unduly surprised (assume that we have heard some sounds before); and, likewise, if someone else were to create the sound, it is usually not too hard to figure out what settings were used. Musicians do this routinely. Sounds, including their timbres, have a rich phenomenological structure that synthesizers manage to reveal very successfully. And once we understand how all this works, we can successfully explain (at least, up to a point) why such sounds sound the way they do.

So what would a colour synthesizer be like? We have all seen the customized colour rectangle on a computer, where the hue-spectrum goes from left to right, and saturation is maximal at the top and zero at the bottom. Lightness is determined by position on a separate scale, and for each result we are told the exact levels of red, green and blue that are used to generate the colour. What we need to do is to take this further so that hues themselves are also parametrized. Thus, in choosing a particular hue, we may set the unique/binary switch at either 1 or 2, set the warm/cool switch at either 1, 0 or −1, perhaps set the 'sharp/unsharp' switch (for want of a better name) likewise at

[13] It is particularly instructive to see how a single note of a square wave (which sounds roughly like a clarinet) can be made to sound like a chord formed from pure sine waves by removing and replacing harmonics. This helps to show why (undetectable) inversions of timbre are much harder to envisage than inversions of hue.

either 1, 0 or −1, and so on. So arranged, the hues depicted in the above Hering colour circle are given the following assignments:

Red	(1, 1, −1)	Green	(1, −1, 1)
Orange	(2, 1, 0)	Turquoise	(2, −1, 0)
Yellow	(1, 1, 1)	Blue	(1, −1, −1)
Chartreuse	(2, 0, 1)	Purple	(2, 0, −1)

Each hue is given a different assignment, which shows that we have enough structure to be getting along with. However, there are $2 \times 3 \times 3 = 18$ possible permutations, and a natural question is what the remaining 10 assignments would look like. Many can probably be ruled out as impossible, but the fact that we are now in a position even to formulate such questions is significant. Of course, the parametrization is still very crude, but could, perhaps, be made more sophisticated. If so, then the possibility is that some alien hues will become describable (though not yet perceivable), including, perhaps, Levine's 'cool red' and 'warm green'. Given the sort of information yielded by this (notably, the systematic inter-modal connections), it may well be that we would be able to tell which is hue is which when we see them for the first time, and just by looking.

All this may sound seriously optimistic, but a *full* explanation of colour *qualia* needs to be seriously ambitious if it is to accommodate alien hues and how things appear to creatures with very different kinds of brain. Thomas Nagel, when discussing panpsychism, talks in this context about a 'mental chemistry', a systematic method of generating indefinitely many possible types of conscious experience from a base set of 'proto-mental properties' possessed by the physical constituents of the conscious subjects in question.[14] *If* such a programme can be made to succeed, then we may perhaps indeed claim finally to have solved the mind–body problem, at least as applied to primitive, non-intentional states. But the sheer scale of the task seems impossible, and we might not unreasonably suspect that we are asking for far too much.

5. Reductive Explanation

The culprit here, I think, is not the parametric model, but a narrowly reductionist conception of explanation. This might sound odd since

[14] Thomas Nagel, 'Panpsychism', in *Mortal Questions*, (Cambridge: Cambridge University Press, 1979), 181–95.

Nagel explicitly rejects reductionism, but what he means is the reduction from the mental to the physical. The move from the mental to the proto-mental presumably counts as reductionist, as the analogy with chemistry implies. Likewise, those with more modest ambitions, such as Hardin and Clark (they are concerned only with human *qualia*), claim explicitly that what they aiming at is a reductive explanation (of *qualia* to the physical). But what is a reductive explanation? In the sense that is relevant here, As are reducible to Bs if and only if the laws governing the behaviour of As are deducible from the laws governing the behaviour of Bs (plus some bridge principles). It is in this sense, for example, that it is widely supposed that chemistry is reducible to physics, or that the behaviour of macrophysical objects is reducible to that of microphysical objects. Given the laws of particle physics, and given information about the microphysical constitution of macrophysical objects (the bridge principles), we may logically deduce the laws governing the behaviour of ordinarily sized objects. The fearsome complexity of macrophysical objects ensures that the actual deduction may be unfeasible, to be sure, but the point is that once the Bs have been fully explained, and we know how the Bs make up the As, there is no more explaining that needs to be done.

But what could a theoretical reduction of *qualia* to the physical amount to? The immediate problem obviously concerns the bridge principles. We do not worry too much about them when reducing chemistry to physics, or the macrophysical to the microphysical, because they do not seem to require any independent explanation. The laws of particle physics will include terms such as 'proton' and 'electron', for example, but not 'carbon'. So, in order to derive the results we want, we must explain what carbon is in terms of atomic particles. We must say, for example, that a carbon atom is one with six protons in its nucleus (and so on), and this would be the relevant bridge principle. Should someone demand an explanation of this, we should simply insist that this is what carbon *is*. This is not to say that the principle is merely analytic, for that would make it *a priori*, but we have the kind of metaphysically necessary identity that simply does not need any explanation, even if a huge amount of empirical effort was needed in order to establish it. If you wonder why carbon atoms have six protons in their nuclei rather than seven, for example, you are, in effect, wondering why carbon is carbon rather than nitrogen: and what kind of question is that? More generally, *explaining* why a whole should have the parts that it does (as opposed to merely ascertaining the fact) is not usually a serious issue. The point is that, once we have established what the constituents of a given

composite whole might be, it is unusual to demand an explanation as to why it should be so. Moreover, if an explanation does seem to be called for, it is surely only because we are puzzled that the constituents should have arranged themselves in that particular way in the first place. But in that case, it is the base theory itself that we need to examine, not the bridge principles.

When it comes to giving a reductive explanation of colour phenomenology to brain functioning, however, things are clearly different. Even if we have a complete theory of how the objectively measurable results of psychophysical testing are explained by neurophysiology, we need bridge principles to ensure that the phenomenological language within which our problem was originally formulated enters the system. We thus need to add to our base theory principles such as

S sees red if and only if […]

where '[…]' uses only the (non-phenomenological) language employed by the base theory. But, it may be protested, such principles are surely *exactly* what we are trying to explain, and exactly what the base theory on its own will *not* explain!

This is perhaps to exaggerate the point, for the base theory may succeed in telling us quite a lot that we need to know. But it remains clear that the really crucial facts are going to be left unaccounted for. It is, again, significant that Nagel talks here of 'mental chemistry'. The relation of a chemical compound to its constitutive elements is of the whole–part kind, and is therefore explanatorily unproblematic. Should we therefore come to know, firstly, what the laws of proto-mental qualities are, and secondly, how each mental quality is composed of its proto-mental constituents, then we shall have achieved our (reductive) explanatory goal – and the bridge principles will not be an issue. This is why panpsychism, if it could ever be made to work, would be such an attractive option – indeed, surely, an inevitable option for a reductionist non-physicalist. Unfortunately, we have absolutely no idea what 'proto-mental' properties are going to be like, and consequently cannot proceed any further in this direction. I prefer instead to talk of parametrization, and this is slightly different, for a parameter is an *aspect*, not a *component*, of what it parametrizes. Although the connection between a *quale* and its parametric settings might seem to be neither more nor less problematic than that between a whole and its parts, the fact remains, however, that the parametrization primarily shows the relations *between qualia*. It does not attempt to provide *reductive* bridge principles to the physical (or to anything else), and is consequently far more likely to lead to explanatory success.

Again, this might seem to be too swift. Austen Clark's elaborately worked out strategy is to identify *qualia* by locating them within a single quality space organized by a similarity metric. Pure phenomenology (supplemented and refined by precise psychophysical measurements) yields the level of similarity between any two elements within the scheme, and multidimensional scaling – a technique used in statistics – is then used to construct the whole quality space.[15] Phenomenological terms are therefore not defined directly – from the base theory or anything else – but indirectly, by means of their internal relations. This is not unlike my approach (though Clark speaks of 'differentiative properties', rather than 'parameters'). However, Clark goes further, and argues that if it could be shown, as is perhaps possible (though surely not inevitable), that the resulting structure contains enough internal asymmetries that it maps uniquely onto the underlying physiology, then we shall obtain a reductive explanation. The bridge principles, which so perplex us, are thus constituted by purely structural features of the system.

This is ingenious, but some problems remain. Firstly, as we have already noted when we discussed diagonal hue-inversion, asymmetries do not themselves automatically yield explanations. This might no longer be the case if we knew *all* the asymmetries, of course; but we need to start explaining things long before we ever get to that stage. Secondly, there is something unsettlingly austere about structuralist analyses. They tend to deliver pure form at the expense of content, and it is the content that we want to know about here. It may be protested that my parametric account is also structuralist, but I do not attempt to *reduce* content to form in the way in which Clark does, so it is unclear that the same objection applies here. Thirdly, and perhaps most significantly, we need to remember that if Clark's program were really to work, then we would have solved what David Chalmers calls the 'hard problem', i.e. the problem of explaining why neurophysiological processes should yield any kind of consciousness whatsoever. A stubborn Cartesian intuition remains, namely that 'zombies' (i.e. unconscious physical duplicates of ourselves) remain, at least, a theoretical possibility. Perhaps this intuition is just wrong; but even it is, it remains very hard to see just why a programme such as Clark's should have managed to prove this. It is certainly useful to be given quality

[15] 'For example, from a map of the United States it is easy to measure inter-city distances. Multidimensional scaling proceeds in the reverse direction: given a table of inter-city distances, it reconstructs the map' (Clark (1993), 210).

connections of this type, but we may still wonder if they are tight enough to ensure reduction. Objective psychophysical investigations of the relevant kind would yield indistinguishable results even if applied to our zombie twins (should there be such things); which suggests that the fundamental assumption that we have consciousness in the first place is not automatically guaranteed, as it would need to be.[16]

These points can be difficult to see, since the current debate focuses primarily on slightly different questions. The status of reductive analyses is much discussed, but primarily in the context of whether reduction requires conceptual analysis. Ned Block and Robert Stalnaker (1999) claim it does not, whereas Chalmers and Frank Jackson (2001) claim that it does; and the discussion concentrates largely on delicate issues about the relationship between metaphysical necessity and the *a priori*.[17] The question of whether explanation requires reduction *at all*, with or without conceptual analysis, therefore tends to be sidelined.[18] The significance of the part–whole relation is likewise under-appreciated. It therefore becomes harder to see the real reason why Clark's explanatory project is unlikely to hit the target: it simply aims too high.

6. Reduction versus Connection

But what alternative model of explanation is there? My suggestion is that Clark is right to search for connections, but that the connections do not need to be all that strong to be explanatorily adequate. Nor do we need an asymmetric grounding in the physical. Rather, what we

[16] See also Chalmers (1996), 235.

[17] Block, Ned and Robert Stalnaker, 'Conceptual analysis, dualism and the explanatory gap', *The Philosophical Review* **108** (1999), 1–46; David Chalmers and Frank Jackson 'Conceptual analysis and reductive explanation', *The Philosophical Review* **110** (2001), 315–60. See also David Carruthers, 'Reductive Explanation and the Explanatory Gap', *Canadian Journal of Philosophy* **34** (2004), 153–74; Ausonio Marras 'Consciousness and Reduction', *British Journal for the Philosophy of Science* **56** (2005) 335–361; Neil Campbell, 'Why We Should Lower Our Expectations about the Explanatory Gap', *Theoria* **75** (2009), 34–51; and Kevin Morris 'Does Functional Reduction Need Bridge Laws? A Response to Marras', *British Journal for the Philosophy of Science*, **60** (2009), 647–657.

[18] However, Steven Horst has usefully challenged many assumptions here, in *Beyond Reduction: Philosophy of Mind and Post-Reductionist Philosophy of Science*, (Oxford: Oxford University Press, 2007).

need is a more democratic relationship between mental and physical elements, and a mutually supportive interlocking structure – a kind of natural harmony that (perhaps) prevailed before Descartes and the mechanistic philosophy introduced a radical and unbridgeable schism.

This may sound fanciful, but what do we actually require of explanations of colour phenomenology? The reason that such-and-such neurophysiological processes correlate with seeing green is hard to fathom because, at first sight, there seems to be no reason why they could not just as easily correlate with seeing red or blue or anything else. But notice that we do not experience similar perplexity when considering, for example, the fact that light of high intensity looks bright. Nobody is likely to wonder whether you might see things growing dimmer when I see them growing brighter.[19] Why is this?

Firstly, we must recognize that there is no strict impossibility here. People often conflate *luminance* (a photometric quantity) and *brightness* (roughly, its psychosensorial correlate), and the terminology here can be confusing. But they are evidently different magnitudes, and their connection is, in fact, quite complex.[20] True, it seems overwhelmingly natural to suppose that a dangerously high level of luminance will result in a painfully bright sensation: we cannot imagine what a painfully dark sensation would be.[21] But this merely relates to what we can readily imagine, not what absolutely has to be. The inverted *qualia* scenario surely remains perfectly *possible*, even if it is one that we are not tempted to worry about.

Other inverted *qualia* scenarios are also hard to imagine in any detail. Is it at all likely that you hear things getting louder where I hear them getting softer? Again, no, and for the same sort of reason. Might you hear a higher pitch where I hear a lower one? This, perhaps, is more easily imaginable but, again, leads to difficulties when we reach extremes. A very low pitch is 'heard' as much through our feet as through our ears, and for straightforward physical reasons. Could a very high pitched sound have that kind of vibratory phenomenology? Again, there is no contradiction here, as far as I can see; but, for all that, people seldom take the possibility seriously.

[19] Though see David Cole (2000), 'Inverted Spectrum Arguments', at http://www.d.umn.edu/~dcole/inverted_spectrum.htm

[20] Levels of luminance provide a prediction of levels of perceived brightness only in certain narrowly defined circumstances, and even there the correlation is logarithmic, not linear.

[21] For a similar reason, black/white inversion (as with a monochrome photographic negative) is not as readily conceivable as it might appear.

Explaining why we hear things normally, and not in reverse-pitch, seems quite unnecessary, since the connection between sound-wave frequency and normally perceived pitch just seems so *natural*: there is no significant itch that needs scratching. Could you taste bitterness where I taste sweetness, and without anyone noticing? Or feel severe pain where I feel a highly pleasurable sensation? The same point again. In fact, what is remarkable (but not always remarked upon) is the sheer peculiarity of *hue* phenomenology: it is, I think, the *only* psychologically convincing example where an undetectable inverted *qualia* hypothesis can be taken seriously.

Why is this? The reason, surely, is that hues do not have the obviously recognizable internal structure to be found in the other examples, and we have already discussed how we might rectify this. The lack of reductive explanation, by contrast, is not the issue at all. We lack reductions with the other examples as well, but do not lack explanations (of course, we do not have *complete* explanations, but we hardly ever get those anywhere). The crucial point is that if we could somehow manage to analyse hues in such a way that inverted hue scenarios come to seem to us just as contrived and implausible as the other inversion scenarios, then we shall have achieved a truly remarkable feat of explanation. Nobody should doubt the enormous significance of such a result. But physicalism does not have to be established (or even assumed) for us to do this.

7. Non-Physicalist Explanations

Not everyone is scared of dualism: for example, Chalmers is happy to go along with it. On his view, we simply need to accept that there are fundamental laws of nature that ensure that when physical systems reach a certain level of complexity, then a certain type of consciousness will be generated. Such laws are ultimately contingent – brute facts, if you will – but then so are all laws of nature, as we have learnt from Hume. That such-and-such brain processes cause a subject to see green may be, in the final analysis, inexplicable; but, then, so is the fact that one billiard ball will get another to move upon impact. We do not lose much sleep over the latter; nor, perhaps, should we over the former.

The point is that there are limits to which anything at all can be fully *explained*. Laws play a crucial role in explanation, but they cannot be eliminated altogether from the *explanantes*. We can explain many laws in terms of more basic ones, of course, but the most basic laws must forever be unexplained – unless they turn out

to be self-explanatory, which is surely very unlikely. We tend to assume that such basic laws will be microphysical, and not include anything psychological; but, as Chalmers says, we have no real reason to think this.

Yet what could Chalmers' ultimate psychophysical laws be like? It is no use just having a law that says, 'Whenever you get a physical structure of general type P, then you get consciousness', for consciousness is not a single uniform quality common to all and only conscious states. You cannot, for example, transform your zombie twin's functional states into the full-blown mental states that you yourself experience by just adding to them a single extra ingredient, 'Consciousness', unless this ingredient is very artificially contrived, and therefore quite out of place in any serious law of nature. Since all types of consciousness, human or otherwise, will eventually have to be accounted for, we must concede that we do not, at present, have the slightest idea what such basic psychophysical laws could be like. The problem is not that we cannot explain why such laws should exist in the first place – I agree with Chalmers, following Hume, that such an explanation is uncalled for. It is rather that we cannot even begin to say what they are going to look like, or in what language they should be couched. They therefore cannot be used in explanations in anything like the sort of way in which, for example, the laws of particle physics can be used to explain the laws of chemistry. But what this shows, once again, is that a different model of explanation is needed, one which is connectionist rather than reductionist. Chalmers agrees with much of this, but still places a residual emphasis on the eventual discovery of these ultimate laws.[22] My proposal, by contrast, suggests ways in which we can make explanatory progress without worrying about such things *at all*. The difficulty, after all, is not just that we cannot reduce colour phenomenology to physics: we cannot even imagine reducing it to *anything* – physical, mental, proto-mental, or whatever! The move from physicalism to dualism will therefore not assist us in the way in which Chalmers thinks. Indeed, the whole issue of physicalism versus dualism may turn out to be a red herring as far as the *explanation* of perceptual qualities are concerned. (Of course, the issue remains highly relevant as far as their *metaphysics* is concerned, but we do not need to attend to all of metaphysics when we attempt to explain things.)

This may still seem implausible, and for several reasons. Firstly, it may still be insisted that if perceptual qualities are not necessitated by

[22] Chalmers (1996), 213–8.

physical qualities, then brain structure cannot possibly help to explain phenomenology: how could it if the same brain structure could equally produce a different phenomenology (or none at all)? This point, which seems crucial to many, does not obviously rely on a bias in favour of reductive models. Secondly, it may also be objected that non-physical magnitudes would inevitably lack the kind of measurable precision that is essential for scientific progress. We might hope (and I do) that if our phenomenal concepts become adequately parametrized and connected in the manner I have indicated, then they will thereby become scientifically respectable in their own right, and regardless of any physical underpinning. But subjective, introspective reports are notoriously unreliable, and it should be remembered that colour vision science only started to make real progress (in the middle of the last century) with the advent of objective methods of psychophysical testing. Thirdly, it may be protested that the emphasis on what is 'natural' and 'intuitive', as opposed to what is 'puzzling', 'contrived' and 'implausible', which plays a leading role in my account of why inverted hues differ from other sensory inversions with regard to what does, and does not, need to be explained, seems to ignore the embarrassing fact that the most obvious and effective way to eliminate puzzlement is simply to reduce curiosity! Such are the perils of psychological criteria of explanatoriness. All in all, it may be feared that my non-physicalist approach to explaining colour phenomenology can amount to nothing more than a great leap backwards.

But these objections are far from decisive. Firstly, the search for neurological correlates of inter-sensory resemblances certainly does not assume that such correlations are metaphysically, and not just causally necessary: it should be remembered that ordinary scientists are quite indifferent to the distinction. Moreover, the claim that if physicalism fails, then it would follow that the same brain structure could equally produce a different phenomenology (and therefore fails to explain the actual phenomenology), is tendentious. We do not usually think that different effects could have been produced from the same causes merely because the underlying laws are not metaphysically necessary. For example, we do not say that the contingency of the inverse square law of gravitation implies that planetary orbits could have been other than they are, and hence that there must be an untraversable explanatory gap somewhere. It depends a bit on what is meant by 'could', of course; but if it is taken to mean 'could equally' or 'could just as easily', then the claim is simply false. All we lack is a kind of idealized super-explanation that might have appealed to Spinoza or Leibniz, but

which few people take seriously nowadays. Scientific explanation should not be assimilated to mathematical proof.

Secondly, we may concede that the use of subjective reports certainly makes life difficult if we aim for scientific precision; but to ignore them altogether is to risk losing sight of our *explanandum*. A colour vision science that establishes many precise, agreed results, but does not attempt to explain why colours actually look the way they do, is clearly impoverished. Moreover, the quantitative results achieved by Hurvich and others which established the opponent-process theory still lean heavily on the original ideas of Hering and the results of simple introspection. It could hardly be otherwise. This earlier, nineteenth-century tradition in physiology, before the subject split into 'a bodiless psychology and a soulless neurology', in Oliver Sacks's words, had better not be dismissed too readily if we wish to explain phenomenology.[23]

Thirdly, it is true of any plausible conception of explanation, not just mine, that there has to be an internal connection between what we need to explain and what we find puzzling: psychological criteria must certainly play a role here.[24] Of course, we need other, non-psychological criteria as well, but my thesis does not suppose otherwise, and I can see nothing in any of my proposals that demands that we wilfully neglect, or attempt to reduce, our intellectual curiosity in any discreditable sense. It is just that we cannot afford to be curious about everything at once. An explanation does not cease to be valid just because it fails to explain everything. On the contrary, good explanations are invariably partial, not complete, since what interests us would otherwise be buried amongst a mass of irrelevant detail.

The mind–body problem appears to be as unsolvable today as ever, but many of the difficulties here are surely self-induced. The closer we bring mental states to their neurological correlates, the stronger our thesis gets and hence the more difficult it is to explain how it could be true. Yet the further apart they are kept, the more difficult it is to explain why they are connected at all. In short, we lose both ways! We could continue to bang our heads against a brick wall, of course, and we probably shall. But it might be better to stand back

[23] Oliver Sacks, *The Man Who Mistook His Wife for a Hat*, (London: Picador, 1986), 88. See also x–xi. His main regret is that clinical descriptions of neurological disorders too often seem to depersonalize the patient. However, the idea may be generalized into a broader critique of Cartesianism and its successors.

[24] See also Campbell (2009).

and to question some of the assumptions that led us to this *impasse*. The correct shape of phenomenological explanations is one thing that could be usefully re-examined.[25]

Lancaster University
N.Unwin@lancaster.ac.uk

[25] Earlier versions of this paper were read at research seminars at Lancaster University and the University of Central Lancashire, and at a Royal Institute of Philosophy seminar at the University of Bradford; and I am grateful for the many useful comments made.