Jeffrey A. Gray, Susan Chopping, Julia Nunn,
David Parslow, Lloyd Gregory, Steve Williams,
Michael J. Brammer and Simon Baron-Cohen

Implications of Synaesthesia for Functionalism

Theory and Experiments

Abstract: *Functionalism offers an account of the relations that hold between behavioural functions, information and neural processing, and conscious experience from which one can draw two inferences: (1) for any discriminable difference between qualia there must be an equivalent discriminable difference in function; and (2) for any discriminable functional difference within a behavioural domain associated with qualia, there must be a discriminable difference between qualia. The phenomenon of coloured hearing synaesthesia (in which individuals see colours when they hear or see words) appears to contradict the second of these inferences. We report data showing that this form of synaesthesia is genuine and probably results from an aberrant projection from cortical language areas to a region (V4/V8) specialized for the perception of colour. Since functionalism purports to be a general account of consciousness, one such negative instance, if it can be further sustained empirically, is sufficient to invalidate it.*

Introduction

Synaesthesia is a condition in which, in otherwise normal individuals, stimulation in one sensory modality reliably elicits the report of a sensation in another. Since synaesthesia has recently been extensively and lucidly reviewed in the

Correspondence:
Jeffrey A. Gray, Institute of Psychiatry, De Crespigny Park, Denmark Hill, London SE5 8AF, UK.

Other authors currently —at same address except:
Julia Nunn, Department of Psychology, Goldsmiths College, London SE14 6NW, UK.
Lloyd Gregory, Section of GI Sciences, University of Manchester, Hope Hospital, Stott Lane, Salford, M6 8HD, UK.
Simon Baron-Cohen, Department of Experimental Psychology, Cambridge University, Cambridge, CB2 3EB, UK.

Journal of Consciousness Studies, **9**, No. 12, 2002, pp. 5–31

pages of this journal (Ramachandran & Hubbard, 2001a) and elsewhere (Grossenbacher & Lovelace, 2001; Rich & Mattingley, 2002), we shall not go over this ground again, except to pick out certain points particularly relevant to our theme. Statements of fact regarding synaesthesia, if not specifically supported by other references, rely upon these reviews. Our experimental work (some of which is reported for the first time below) has been concerned exclusively with one kind of synaesthesia, namely, word–colour synaesthesia. In this, heard or seen words elicit (besides the auditory or visual percepts normally elicited by such stimuli) sensations of colour. This is a common form of synaesthesia. Most of the subjects described by Ramachandran and Hubbbard (2001a,b) were drawn from subjects with a condition similar to word–colour synaesthesia, that is, grapheme–colour synaesthesia, in which the symbols for numerals evoke experiences of colour. In what follows, we assume that conclusions based upon their sample of subjects and upon our own are mutually complementary. Although we do not discuss other forms of synaesthesia, we know of no reason why the conclusions we draw would not apply to synaesthesia generally.

Our concern in this paper is to draw out the implications of word–colour synaesthesia for the dominant approach to the problem of consciousness in contemporary science and philosophy alike, namely, functionalism. Since functionalism purports to provide an entirely general account of consciousness, then any phenomena that cannot be fitted into its conceptual framework may seriously damage it. We try to show here that there are certain facts about synaesthesia which are very difficult to fit into the framework of functionalist thought (for earlier presentations of some of these arguments, see Gray *et al.*, 1997; Gray, 1999).

Functionalism

It is not our intention to give a detailed scholarly analysis of functionalism (see Dennett, 1991). We shall rather state this doctrine in a form in which, we believe, it is imperilled by our empirical findings, as outlined below. It may turn out that variants of functionalism can be proposed that will address some of the issues we raise in this paper, but in that case we shall at least have sharpened up what a successful functionalism must claim if it is to account for synaesthesia. Moreover, it is possible that variants of our arguments and/or variants of our experiments will be able to address those variants of functionalism. Most importantly, since our case rests on empirical data as well as theoretical arguments, we shall have taken a small step towards bringing these issues into the experimental laboratory — which is where we think they belong.

The crux of the 'Hard Problem' of consciousness lies in the phenomena of perception — qualia. Consider, as a specific version of the Hard Problem, this question: how should one explain the difference between two subjective experiences of colour, say of red and green? We take functionalism to approach a question of this kind in the following way.

Functionalism starts by eliminating from the question the qualia of red and green as such. For these it substitutes as the explicandum the repertoire of behavioural responses by which the experiencing individual demonstrates the capacity to discriminate between red and green. This repertoire would include e.g. pointing to a red (green) colour when requested to do so, using the word 'red' ('green') appropriately in relation to the colours red and green, stopping (going) at red (green) traffic lights, stating that a lime is 'green' and a tomato 'red', and so on. Next, functionalism seeks an understanding of the mechanisms by which these behavioural 'functions' are discharged. This understanding may be sought at a 'black-box' level, as in the familiar box-and-arrow diagrams of cognitive psychology, neural networks, computer simulations and so on; or it may be sought in the circuitry of the actual brain systems which connect the inputs to the outputs of each of the discriminating behavioural functions. A full 'function' for a given difference between qualia then consists in a detailed account of the corresponding differences in inputs, in outputs, and in the mechanisms that mediate between input and output. If such a full functional account is given, then, according to functionalism, there is no further answer that can be given to the original question: what is the difference between the subjective experiences (the qualia) of red and green? To continue asking this question in the face of such a complete functionalist account would be a meaningless activity. For, according to functionalism, qualia just are the functions (input–mechanism–output) by which they are supported.

From this formulation of functionalism one can draw the following, 'primary', inference: for any discriminable difference between qualia there must be an equivalent discriminable difference in function. We consider here also a 'complementary' inference: for any discriminable functional difference, there must be a discriminable difference between qualia. Clearly, there are ways in which this complementary inference may be false. There are many forms of behaviour which are not accompanied by qualia at all. So, for example, the pupils of one's eyes constrict if illumination increases and dilate if it decreases; but one is not normally aware of either of these changes in pupil size. However, in the case of a behavioural domain which is normally accompanied by qualia, it seems reasonable that, whenever functionalism draws the primary inference, it should also draw the complementary one. Let us return to the example of red and green. The primary inference is that (within the domain of colour vision), if someone claims to have different red and green experiences, then there must be different functions (input–mechanism–output) to support this claim. The complementary inference would be that (within the domain of colour vision), if someone manifests different functions (input–mechanism–output), then there must be different qualia accompanying them. The two inferences together constitute a claim for identity between qualia and functions within the domain of colour vision. Functionalism at its strongest generalizes this identity claim for all domains and all qualia.

There is a related but nonetheless separate strand of functionalist thought. This treats the functions that give rise to qualia as providing benefit to the behaving organism. This strand is particularly evident in discussions of the evolution of qualia. The claim here is that evolution works by the selection of behavioural functions that contribute (in the usual way) to Darwinian survival and thus of the neural mechanisms which mediate those functions; the evolution of qualia themselves occurs only parasitically by linkage to such functions (Harnad, in press). If this view is correct, one would not expect to find qualia which adversely compete with the functions to which they are linked.

A final word, for our here limited purposes, about functionalism is this. Functionalism is proposed in two different flavours (we say 'flavours' rather than 'forms', since the nuances are often quite subtle). In one flavour, qualia are reduced to so little beyond the functions with which they are linked as to be virtually eliminated (Dennett, 1991). In the other, the separate existence of qualia is explicitly acknowledged, but all empirical data are treated as requiring explanation in terms of the functions with which they are linked. Harnad (in press), for example, argues that qualia are epiphenomena, caused by functions and their underlying mechanisms but having no causal effects of their own. Denial of this position, he believes, entails the espousal of dualism.

In this paper we report evidence from word–colour synaesthesia which appears (1) to contradict what we have called above 'the complementary inference' by demonstrating disparate functions sharing the same qualia; and (2) to demonstrate qualia with behavioural effects that are adverse to the functions with which they are linked. We believe that this evidence constitutes a serious challenge to functionalism in both its flavours.

Nature of the Synaesthetic Experience

We defined 'synaesthesia' at the outset of this paper conservatively: as a condition in which 'stimulation in one sensory modality reliably elicits the report of a sensation in another'. However, if the arguments advanced here are to hold, it must be the case that the report is veridical, in two senses. First, the report must be more than mere confabulation — there must be some experience that is separate from the report and is reliably reported. Second, that experience must be perceptual, if we are to base arguments about qualia upon it.

Over the last decade evidence has accumulated to support these assumptions (Ramachandran & Hubbard, 2001a).

Baron-Cohen *et al.* (1993) demonstrated the reliability of reports of word–colour synaesthesia: their subjects gave essentially identical reports of their colour experiences in response to a list of words at a year's interval, with no warning that they would be retested at that time. The similarity of reported word–colour associations in a group of non-synaesthete controls over a period of just a month was strikingly inferior.

The perceptual nature of the synaesthetic experience is well documented in the Ramachandran and Hubbard (2001a) review. We give here just one example of their experiments (for references, see their paper). It is characteristic of visual perception that items in a display which differ in a feature from other 'background' items in the same display 'pop out' — that is to say, they are seen automatically and involuntarily as being different from the background items, and they are grouped together.as separate from the background items. Ramachandran and Hubbard presented subjects with a black-against-white display of '2's and '5's, computer-generated so that the latter were mirror images of the former. The 5s were disposed among the background 2s so as to form a triangle. Non-synaesthetes found it hard to detect the triangle. Number–colour synaesthetes, in contrast, for whom 2s and 5s elicited different colour sensations (e.g., red and green), at once saw the triangle, which stood out in one colour against a background of a different colour. It is virtually impossible to account for this and similar phenomena except by giving credit to the synaesthetes' own reports of an experienced colour perception.

This evidence from psychophysical experiments is supported by neuroimaging data. In a first experiment of this kind, using positron emission tomography (PET) as the imaging technique (Paulesu *et al*., 1995), a list of spoken words was presented to a group of word–colour synaesthetes and to non-synaesthete controls. The synaesthetes but not the controls showed in response to the spoken words activity that was located in the visual system. In this experiment, the activity was observed at relatively high levels of the visual system, in visual association cortex, a finding that has been interpreted as indicating that synaesthetic experiences of colour 'result from partial activation of higher-order visual cortical networks, rather than arising at the earliest levels of cortical visual processing' (Grossenbacher & Lovelace, 2001, p. 38).

However, we have recently re-investigated this issue (Nunn *et al*., 2002), using essentially the same materials and procedure as Paulesu *et al.* (1995), but taking advantage of the greater temporal and spatial resolution afforded by functional magnetic resonance imaging (fMRI; Buxton, 2002). We measured in both word–colour synaesthetes and non-synaesthete controls, matched for sex (all female), handedness (all right-handed except one subject in each group), and verbal intelligence, regional blood oxygen level dependent (BOLD) activity in response to a list of spoken words (relative to auditory stimulation by tones as a baseline condition). We related the resulting pattern of BOLD activity to that obtained in a standard test of brain activation by colour. In the latter, regional BOLD activity is measured in response to presentations of 'Mondrians' (patterns made up of irregular rectangles, each of a different colour, like those painted by the artist Piet Mondrian) compared to a baseline condition of the same patterns in black and white. We performed two experiments, one in which the pattern of BOLD activation by coloured Mondrians was derived from published data (Howard *et al*., 1998) on non-synaesthetes, and a second in which we re-measured this pattern both in a further group of non-synaesthete controls and in the synaesthete subjects themselves. The results of the two experiments were essentially identical. They demonstrated that the synaesthetes, but not the controls, activated the visual system in response to spoken words, confirming in this broad respect Paulesu *et al.* (1995). However, as illustrated in Figure 1 [see back cover], the activation was located lower down in the visual system, exactly in the region known from earlier reports (and replicated in our own Mondrian experiment) to be the earliest point at which the brain computes colours as such: viz, areas in the fusiform gyrus known as V4 or V8 and as V4 (Bartels & Zeki, 2000; Hadjikhani *et al*., 1998).[1] Thus these findings, like those of Ramachandran and Hubbard's (2001a,b) psychophysical experiments, support the hypothesis that synaesthetic colour experiences arise at an early stage of visual processing and are truly perceptual in nature.

The synaesthetic experience, then, at least in the case of word– or number–colour synaesthesia, is reliable, veridically reported and perceptual. This experience, in addition, provides a particularly uncluttered example of the general truth (Velmans, 2000) that perceptual experiences are constructed by the brain, and are only rather indirectly related to the states of affairs in the 'real world' that cause them to be constructed. The perceptual experiences of colour that, in word–colour synaesthetes, are reliably, automatically and involuntarily elicited by words bear no relationship at the time of their elicitation (we consider below whether they may have borne such a relationship at some earlier time) to the wavelength properties of light reflected from surfaces which normally provide the external basis for experienced colour. There can be no question, therefore, that synaesthetic colours are constructs of the brain. They leave no room for interpretation within a 'naïve' or 'direct' perceptual realist framework (e.g., Searle, 1983, p. 57). For such an interpretation even to get off the ground, there has to be some kind of resemblance or correspondence between the state of affairs in the external world that gives rise to the percept and the percept itself (though this begs the question of what is meant by 'resemblance' in this context). No such resemblance exists when a synaesthete reacts, say, to the heard word 'train' with a greenish-blue percept. When qualia of this kind are experienced, therefore, they cannot be regarded as constituting direct perception of any state of affairs in the real world.

As an aside that could take us into a wider discussion than we wish to enter here, we note that Wager (1999) has used similar arguments, also based upon synaesthesia, in an effective critique of representationism, i.e., the view 'that the phenomenal character of an experience supervenes on its representational content'. And a recent and influential neo-behaviourist account of perception by O'Regan and Noë (2001), is also open to the same line of attack. This account treats different qualia as reflecting the different sensorimotor contingencies that govern their occurrence. But red colours experienced by a coloured hearing synaesthete via the visual or auditory systems, respectively, participate in radically different sensorimotor pathways yet are perceptually closely similar; while the colour green accessed via the visual route resembles the same colour accessed by the auditory route, rather than the colour red governed by the same visual sensorimotor contingencies.

[1] The nomenclature V1, V2 etc is used to designate cortical areas of the visual system. Increasing numbers provide a (rough) indication of increasing level in the system. Each area, furthermore, is specialized to deal with different aspects of visual input, e.g., V4 for colour, V5 for motion (Zeki, 1993).

Function vs Tissue

We have so far presented a formulation of functionalism without contrasting it to any alternative approach to the understanding of qualia. In the context of this discussion of synaesthesia, the most relevant contrast is with what we shall call, for want of a better word, a 'tissue' approach. The 'want of a better word' reflects the fact that this alternative to functionalism has been articulated far less clearly than functionalism itself. Indeed, it is not clear that it has ever been articulated at all.

What *has* been clearly articulated, and at one time very forcefully argued, is the 'mind–brain identity' theory (Borst, 1970). This holds that mental states (conscious or otherwise) are identical to brain states, and that a given mental state will be fully accounted for if and when one has accounted for the corresponding brain state. This theory eliminates qualia as completely as does functionalism. It may, of course, turn out to be correct. But to assume so in advance prevents one from putting questions of major scientific interest (Gray, 1971; 1987). So, for example, if qualia are identical to brain states, one cannot even ask (what appears to be a valid scientific question), 'how does the brain create qualia?' A 'tissue' approach would allow this question to be put, and would anticipate an answer of the general form: brain processes of type X by way of causal pathways P give rise to qualia of type Q. A position of this kind, but stated only at a very general level, has been defended by Searle (1987), in his proposal that mental phenomena arise from 'macro-properties' of the brain. Note, however, that in such a formulation a 'brain process' taking place in 'brain tissue' may be critical for conscious experience in virtue of just those particular kinds of *neural* function carried out there, and for which the tissue is in some way specialized. But this is a sense of function that is very different from the behavioural input-output sense with which we here contrast the 'tissue' approach.

Functionalism more or less inevitably leads to the conclusion that, if a system displays behaviour of a kind that, in humans, is associated with conscious experience, then the components out of which the system is made are irrelevant (as one among many possible examples, see Aleksander, 2000). The contrary, 'tissue' view is that there is something special about the physical components out of which brains are made that provides a necessary condition for consciousness to arise. This view may stress the physics of these components, as in Hameroff and Penrose's (1996) quantum gravitational theory of consciousness, or their biology, as in Koch and Crick's (2001) search for genes that may underlie the evolution of the neural correlates of consciousness. Views of this kind are sometimes explicitly proposed as superior to functionalism, as by Hameroff and Penrose, but more often it is left unclear whether they are or are not incompatible with functionalism. Similarly, on the functionalist side, some thinkers (see, e.g., Harnad's, in press, discussion of level T4 of the Turing test) concede the possibility that, for a complete account of consciousness, the actual mechanisms that the brain utilizes may be a crucial addition to its functions, whereas others scorn the whole idea as calling upon 'wonder' tissue (Dennett, 1991).

Despite its lack of conceptual articulation, we shall here use as our contrast to functionalism the tissue approach. This choice is dictated by the (rough) parallel this contrast affords to the two most plausible accounts of the aetiology of synaesthesia. These hold that synaesthesia is based upon either (1) early and strong associative learning or (2) an unusual form of 'hard wiring' in the synaesthete brain. The parallel we draw recognizes equivalences between, on the one hand, Hypothesis (1) and functionalism and, on the other, Hypothesis (2) and the tissue approach. More specifically, Hypothesis (2) holds that, as the result of an extra, hard-wired projection in the synaesthete brain, activity in the inducing sensory pathway (e.g., in word–colour synaesthesia, auditory speech analysers) leads automatically to further activity in the induced pathway (subserving colour vision); and that activity in the induced pathway is sufficient — without regard to the nature of any associated ongoing functions — to trigger the associated conscious experience (of colour).

We describe here experiments which aimed to test the first, associative learning account of word–colour synaesthesia. We present the experiments in the next section, before returning to consider their possible wider implications for functionalism in relation to the general problem of conscious perception.

Experiments on Associative Learning as a Possible Basis for Synaesthesia

Synaesthetes generally report that they have had their synaesthesia for as far back as they can remember. They do not normally report any specific learning experience that might have led to their associating a particular word with a particular colour. However, such learning may have taken place at a sufficiently early age to fall into the period of infantile amnesia. Thus, one possible explanation for synaesthesia is that the individuals concerned formed exceptionally strong and enduring associations between words and colours at an early age. Ramachandran and Hubbard (2001a) give a number of reasons against acceptance of this hypothesis, but none that conclusively rule it out.

The alternative account is that the synaesthete brain has abnormal projections that link one part of the brain (the sensory system in which the inducing stimulus is processed) to another (the sensory system in which the synaesthetic percept is experienced). So, in the instance of word–colour synaesthesia, there would be a projection, not existing in the non-synaesthete brain (nor even in the brains of individuals with other types of synaesthesia), from the parts of the brain which process heard and/or seen words to the colour-selective regions of the visual system. This abnormal projection might arise because the synaesthete has a genetic mutation which promotes its growth or a mutation which prevents its being 'pruned' during early development, since at this time non-synaesthete brains too show an abundance of connections that are no longer present in the adult brain. The likelihood of a genetic basis for synaesthesia is strengthened by the fact that there is a strong tendency for the condition to run in families, and especially in the female line. These possibilities are well and extensively discussed by

Ramachandran and Hubbard (2001a; see also Harrison & Baron-Cohen, 1997; Marks, 1997).

There is at present no way directly to test the hard-wiring hypothesis, since this would require anatomical investigation of the brain. What we have tried to do, therefore, is to test the associative learning hypothesis. We performed two experiments. For details of the fMRI methods we employed, see Nunn *et al.* (2002).

Experiment 1: Word–colour associations in non-synaesthetes

In the first experiment (described in Nunn *et al*., 2002, as Experiment 3), we trained non-synaesthetes outside the scanner on a series of word–colour associations and then tested them with fMRI to see whether their pattern of BOLD activity in response to spoken words had come to resemble the pattern spontaneously displayed by synaesthetes (as shown in Figure 1 [see back cover]). We made strenuous efforts to ensure that the subjects had formed strong associations between the words and the colours. First, we gave them extensive over-training. We were concerned that, nonetheless, the contextual shift from the training environment to the scanner would weaken these associations. We therefore retrained the subjects once they were in the scanner. Finally, since the synaesthete experience is clearly perceptual, we asked our subjects to 'imagine' the colour associated with each word, and also included as a comparison a condition in which they were asked only to 'predict' the colour. We anticipated that, if the associative learning hypothesis of synaesthesia is correct, then these non-synaesthete subjects should show, particularly in the 'imagine' condition, at least some activation in the V4/V8 region in which the synaesthetes showed activation in response to the same word list

Details of our training and testing methods were as follows. The non-synaesthete subjects learned eight word–colour pairings based upon those described by the synaesthetes. Words were chosen from the list used to display V4/V8 activation in the latter subjects (Figure 1 [see back cover]), with the constraints that each triggered a different colour and that a maximum of one word–colour correspondence was derived from each synesthete. Subjects sat in front of a desk-top computer showing eight colours in a 2 x 4 grid. Clicking on a colour caused a word to be presented through headphones and the computer screen to fill with the colour. This colour remained on the screen until another was chosen. To test learning, subjects heard single words, initiated by themselves, through headphones, in random order, and clicked on the colour paired with that word. No feedback was given. Subjects had to be 100 % correct five consecutive times. Further cycles of learning and testing were applied until this criterion was reached. For testing in the scanner subjects listened with eyes closed to 30-second blocks of single spoken words alternating with blocks of single pure tones (exactly as in the experiment with synaesthetes), one every three seconds for five minutes. Each subject was scanned using fMRI twice: (1) with instructions to 'predict' the colour associated with the word; (2) with instructions to 'imagine' it, in that order. For 'predict', subjects were required simply to think of the name of the colour associated with the presented word. For 'imagine', they were required to visualize the colour as it had appeared on the computer screen. At the end of each five-minute scan subjects reported the percentage of successful predictions/images. All reported 80–100% success. There was then a relearning phase in the scanner. The words were presented followed immediately by the appropriate colour, back-projected onto a translucent screen over the end of the scanner bore. A short re-test session followed, in which the words were presented one at a time, and subjects responded by naming the associated colour. Subjects were required to be 100% correct. All subjects reached criterion on the first re-test session. Re-learning lasted approximately 2 minutes. Subjects were finally scanned twice more in the 'predict' and 'imagine conditions', as described above.

In sum, four sets of activation patterns to words were gathered from these non-synaesthete subjects. Two were gathered prior to retraining: 'pre-predict' (with instructions to predict the associated colours) and 'pre-imagine' (with instructions to imagine them). Two further sets were gathered after retraining in the scanner: 'post-predict' and 'post-imagine'. Since the aim of the experiment was to determine whether imagining colours might, in normal subjects, activate the same colour-selective region activated in coloured hearing synaesthesia, we analysed the region of interest[2] (ROI) established in the original experiments with synaesthetes as the 10-voxel overlap between the colour region in normal subjects (Howard *et al*., 1998) and the response to heard words in synaesthetes (see Figure 1 [back cover]).

We performed four sets of comparisons (pre-imagine vs pre-predict, post-predict vs post-imagine, pre-imagine vs post-imagine and pre-predict vs post-predict) by analysis of variance at a voxel-wise p value of 0.05 and examined significant differences within the ROI. The total number of observed differences was 1. As 40 tests were carried out (10 voxels x 4 experiments) the expectation of false positives at $p = 0.05$ would have been two.[3] Thus we conclude that there are

[2] The rationale for restricting our analyses to area V4/V8 is that this was the only region of overlap observed in our first experiment (Experiment 1 in Nunn *et al*., 2002) between the fMRI activation pattern elicited in normal subjects by seen colours and in coloured hearing synaesthetes by spoken words. This observation justifies the hypothesis that activity in V4/V8 is the closest neural correlate of colour experience in both cases. The rest of the experiments then set out to test this hypothesis by using the area of overlap as the region of interest for analysis in the subsequent fMRI experiments. We believe that this kind of focussed hypothesis testing is a better use of fMRI than the more common shotgun approach.

[3] We performed an analysis of variance, constrained to the ROI. Spatial autocorrelation is a problematic issue in fMRI. In standard parametric testing using say *t* tests, the critical value for the test would need to be adjusted for this inter-voxel interdependence. (A 'voxel' is the three-dimensional equivalent of a two-dimensional 'pixel', that is, it is the smallest volumetric unit of analysis in an fMRI data set.) To cope with this problem we generate a null distribution of our statistic of interest by randomly selecting samples of the observed size from the pooled group data. This is repeated many times at each voxel using the same order of random numbers and preserving the spatial relationship between the voxels. The output from this is a null distribution that subsumes the inter-voxel correlations. We then set the critical value for the statistic of interest at any desired type I error rate by reference to this distribution (Bullmore *et al*., 2001, and earlier references cited therein).

no significant differences. To account for mapping errors, an area of 5 mm surrounding the mask was included in the comparisons, increasing the number of voxels tested and so strengthening this result. All four data sets were next combined, yielding one large group map. The overlap between this map and the ROI was just one voxel, with a chance expectation of 0-1 voxels (10 voxels tested at p = 0.05). Thus there was no evidence in these comparisons that non-synaesthetes activate V4/V8 under any of the conditions tested. These negative results did not represent any general failure of activation, as might happen for example if the subjects simply did not attend to the stimuli, since there was clear activation in the auditory cortex and regions of the brain concerned with language, such as Broca's area, presumably reflecting the active processing of heard words.

The results of Experiment 1 weaken the possibility that synaesthetic colour experiences are the result of normal associative learning that has led to particularly strong associations. If that were so, the non-synaesthetes given over-training on word–colour associations would have been expected to show at least some activation, when hearing words, in the V4/V8 region.

Experiment 2:
Melody–colour associations in synaesthetes and non-synaesthetes

There is, however, an alternative associationist account of our data. Conceivably, synaesthetes differ from non-synaesthetes in the nature of their associative learning process. Perhaps this is unusually strong. If so, one might more easily train them than non-synaesthetes on a novel association. To test this possibility we used training methods similar to those used before, but for word–colour associations we substituted melody–colour associations; and we trained both word– colour synaesthetes (who however reported no colour experiences in response to music) and non-synaesthetes before testing them, as before, in the MRI scanner. If synaesthetes have generally strong associative learning processes, then they would be expected to show responses to melodies after training in the same V4/V8 region which is activated by words in these subjects.

The fMRI methods in this experiment were as described by Nunn *et al.* (2002). The training and testing methods were essentially the same as in Experiment 1, except that six melody–colour pairings were learned by the subjects, rather than eight word–colour pairings. We reduced the number of paired associates because melody–colour associations were harder to learn than word–colour associations. The melodies were chosen from classical works, e.g., Chopin, Mozart. The duration of each melody, played through head-phones, was four seconds, with the onset of the associated colour, filling a computer screen observed by the subject, occurring half a second after the onset of the melody. During the re-learning phase in the scanner, all subjects reached criterion on the first test session. The same four conditions were studied as in Experiment 1: pre-predict, pre-imagine, post-predict and post-imagine. The subjects of the experiment were the same synaesthetes and controls as contributed both the original findings shown in Figure 1 and the negative results for the controls obtained in Experiment 2.

The results of the experiment showed no significant differences in activation patterns between the synaesthetes and controls; and in neither case was there significant activation in the V4/V8 region activated in the synaesthetes by heard words (as illustrated in Figure 1 [see back cover]). As in Experiment 1, there was however clear activation in the auditory system, so the lack of activation in the visual system could not be attributed to a failure to attend to the stimuli. Thus these results lend no support to the hypothesis that synaesthetes might show particularly effective associative learning. In addition, they clearly distinguish between the brain activation patterns elicited by the kind of sensory association that the synaesthetes spontaneously reported (word–colour) and the kind they denied (music–colour).

Function vs Tissue Revisited

It is always difficult to reject a hypothesis on the basis of negative findings alone. Clearly, we cannot rule out the possibility that, despite the considerable effort we put into over-training non-synaesthetes on word–colour associations (Experiment 1) or both synaesthetes and non-synaesthetes on melody–colour associations (Experiment 2), we were unable to achieve the strength of the early learning which hypothetically underlies word–colour associations in synaesthesia. Perhaps there is something special about the period of early learning which cannot be duplicated in adult subjects. Nonetheless, the complete absence in these experiments of any activation in the colour-selective regions of the visual system except in the case of spontaneous synaesthete word–colour associations casts considerable doubt on the hypothesis that the latter are the fruit of normal associative learning.

Given this (albeit weak) conclusion, we are left by default with the hard wiring hypothesis. We should note, however, that the hard wiring and associative learning hypotheses are not necessarily in total opposition to one another. For one possibility is that early in development there is an unusually strong process of associative learning, but that this strength consists precisely in the formation of permanent hard wiring. This possibility need not concern us further here since, for the remaining portion of the present argument, it is sufficient to suppose that adult synaesthetic perceptual experience arises because of obligatory transmission of neural excitation from one pathway (the inducing pathway) to another (the induced pathway) to which the inducing pathway is abnormally connected (Grossenbach & Lovelace, 2001; Ramachandran & Hubbard, 2001a, b). It is not relevant to the argument whether the hard wiring that underlies this abnormal obligatory transmission is genetically determined or arises because of early learning (although the former possibility is much more likely).

In word–colour synaesthesia the behavioural evidence suggests that the inducing pathway most likely consists in regions in which the auditory and visual representations of phonemes and graphemes are located (Ramachandran & Hubbard, 2001a). Our fMRI data (Nunn *et al.*, 2002) do not directly throw further light upon the inducing pathway, nor would they be expected to do so, since this pathway is presumably activated to a similar degree in synaesthetes or non-synaesthetes, respectively, presented with words. They do, however, sharpen up hypotheses concerning the likely route from the inducing to the induced pathway.

The word–colour synaesthetes in our experiments responded to spoken words by activating the colour-selective region of the visual system *without* activation at any earlier point in the visual pathways, such as V1 or V2, although these regions are activated in normal subjects presented with coloured visual stimuli (Howard *et al*., 1998). This result is consistent with the view advanced by Crick and Koch (1995) that activity in V1 is not sufficient for visual awareness, but goes further by suggesting that such activity is not even *necessary* for the conscious experience of colour. A contrary result — activation of V1 by spoken words — has been reported by Aleman *et al*. (2001) in a single word–colour synaesthete. However, given the size of the group (10) we studied, it is reasonable to conclude that V1 activation by spoken words in word–colour synaesthesia is rare or weak. This pattern of results — similar activation in more central parts of the visual pathway, but V1/V2 more clearly activated by the more 'normal' route of stimulation — has been reported also in studies of colour after-images (Hadjikhani *et al*., 1998), motion after-effects (Tootell *et al*., 1995) and illusory motion (Zeki *et al*., 1993). In contrast, as observed in Experiments 1 and 2 here and also by Howard *et al.* (1998), *imagining* colours is *insufficient* to activate either of these regions, V1/V2 or V4/V8. These contrasting patterns of activation are consistent with the common introspection that after-images and after-effects are true visual percepts, whereas merely imagined visual features are not. However, despite this congruence with introspection, the lack of activation of V1 in these experiments must be treated with caution. There are many possible methodological factors that can lead to a spurious null result in fMRI. For example, the size of V1 can vary by a factor of two in different subjects, so that a failure to observe early visual area activity may result from inter-subject averaging, not an actual absence of activity (V.S. Ramachandran, personal communication).

Nonetheless, overall these results suggest, in agreement with Bartels and Zeki (2000), that activation of modules in the visual system specialized for the analysis of particular visual features, such as colour or motion, is both necessary and sufficient (not requiring supplementation by activity in regions earlier in the visual pathway) for the conscious experience of that visual feature. This generalization is supported also by data on the orientation-contingent colour after-effect (the McCollough effect; Barnes *et al*., 1999) and hallucinatory experiences of colour in the Charles Bonnet syndrome (ffytche *et al*., 1998), in which V4/V8 activation again accompanied the illusory experience. From this point of view, then, word–colour synaesthesia can be viewed as an example of illusory experience in which the triggering stimuli (words) occur with very high frequency, as compared to triggers for other illusions, e.g. colour after-images or motion after-effects, which occur with much lower frequency. In all these cases, however, once the relevant visual module (V4/V8 for colour, V5 for motion) is activated (provided the activation reaches a sufficient degree of intensity; Moutoussis and Zeki, 2002; ffytche, 2002), the illusory experience occurs automatically.

This hypothesis, however, requires a *caveat*. Nunn *et al.*'s (2002) critical comparison was between the data from synaesthetes and the activation patterns elicited in non-synaesthetes in the visual system by colour vs black-and-white. This method would capture down-stream effects on synaesthete V4/V8 activation by words emanating, e.g., from prefrontal cortex, a region often suggested to play a critical role in conscious processing (e.g., Crick & Koch, 1995; Dehaene *et al.*, 1998). Thus our findings suggest that V4/V8 activation does not require additional activation lower in the visual pathways to enter consciousness, but they do not speak to the possibility that supplementation is required by feedback from higher processing systems.

A second *caveat* is that the patterns of activation elicited by spoken words and seen colours were not identical in all respects — far from it. Other brain areas were activated by synesthetic but not visually detected colours (e.g. anterior fusiform gyrus, frontal cortex; Table 1 in Nunn *et al.*, 2002); and other brain areas are activated by visually detected but not synesthetic colours. (e.g. Fig 1 in Nunn *et al.*, 2002). A defender of functionalism could therefore argue that our experiments merely show a trivial overlap between the patterns of brain activation elicited by two different kinds of sensory processing. So, for example, if you were asked to recall the smell of camembert and the taste of sugar, there might well be overlap in the associated patterns of fMRI activation in those regions of the brain concerned with *recalling* qualia; but one would not argue from that there is overlap in the brain processes underlying the disparate qualia of smells and tastes themselves. However, an objection along these lines would miss a key assumption in our argument: namely, that the experience of colour elicited in the synaesthete by either spoken words or seen colours is in both cases due to activity in V4/V8 (along the lines indicated in the previous paragraph). Thus the additional and disparate activation patterns elicited by spoken words and seen colours are irrelevant to the argument we present. The existence of such disparate additional regions of activation is entirely to be expected, given that access to the neural correlate of colour experience (*ex hypothesi*, activity in V4/V8) is gained by way of different routes — the auditory and visual pathways, respectively.

A further reason to expect such disparate additional regions of activation lies in the different emotional responses elicited by synaesthetic and visually perceived colours, respectively. In this connection, Dennett (1991, pp. 383–98) has emphasized that, for a full functionalist account of the difference between responding to e.g. red and green, as outlined above in the section entitled *Functionalism*, it is important to include as part of the relevant discriminating behavioural repertoire all affective and emotional responses, expressed or betrayed preferences, and susceptibility to memory and attentional effects. It is consistent with this emphasis that, for example, Nunn *et al.* (2002) observed in response to heard words in synaesthetes, but not controls, activation in the left posterior cingulate cortex which they attributed to the role played by this region in emotional memory. But additional activity of this 'responsive' type does not weaken our hypothesis any more than does the additional activity on the auditory input side, also observed in synaesthetes only. For there is no reason to suppose that either type of additional activity underlies the specifically chromatic nature of the synaeshetic colour experience. That experience, we believe, is anchored in the activity observed in V4/V8.

A further important aspect of our findings (Nunn *et al*., 2002) is that we saw activation in the word–colour synaesthetes presented with spoken words only in *left* V4/V8. Given the left lateralization of cortical language systems, this left lateralized activation may relate to the fact that it is speech sounds rather than sounds in general which elicit the synesthete's colour experiences. Both the lack of activation in V1/V2 and the left lateralization of the activation in V4/V8 were observed (Nunn *et al*., 2002) in two independent experiments, so these appear to be robust findings. Also consistent with these results, in their PET study of coloured hearing synaesthetes, Paulesu *et al*. (1995) found subthreshold activation of left, but not right V4/V8. Thus the abnormal projection which hypothetically underlies word–colour synaesthesia appears to travel from left-lateralized cortical language systems directly (without involvement of regions lower in the visual system; see above) to left V4/V8. This conclusion is in good agreement with inferences drawn from other data by Ramachandran and Hubbard (2001a). It is not yet possible, however, to say whether this left lateralization tells us anything special about synaesthesia or merely about the language systems which act as the inducing pathway for the particular kinds of synaesthesia studied by both our group and theirs.

A final result from our fMRI study of word–colour synaesthetes deserves mention. The data from the experiment using fMRI to investigate the brain region activated by colour (Nunn *et al*., 2002, Experiment 2) showed good agreement as between synaesthetes and non-synaesthete controls — but only in the *right* hemisphere. In this hemisphere, both groups showed activation of V4/V8. However, left V4/V8 was activated by coloured (vs black-and-white) Mondrians only in non-synaesthetes. Thus, in the synaesthetes, left (but not right) V4/V8 was activated by spoken words and right (but not left) V4/V8, by coloured Mondrians. These data raise the interesting possibility that, in word–colour synaesthesia, the putative abnormal projection from left cortical language systems to left V4/V8 prevents the normal dedication of the latter region (together with its right-sided homologue) to colour vision.

Taken together, these results and our inferences from them paint the following picture. Word–colour synaesthetes are endowed with an abnormal extra projection from left-lateralized cortical language systems to the colour-selective region (V4/V8) of the visual system, also on the left. Whenever the synaesthete hears or sees a word, this extra projection leads automatically to activation of the colour-selective region. Activation of this region is sufficient to cause a conscious colour experience, the exact nature of that experience depending upon the particular set of V4/V8 neurons activated. Importantly, there is *no* evidence that the experienced colour plays any functional role in the synaesthete's auditory or visual processing of words. (In the next section, indeed, we report evidence that the experienced colour may actively interfere with such processing.) *Thus, there is no relationship between the occurrence of the synaesthete's colour experiences and the linguistic function that triggers them.* This conclusion appears to be incompatible with the functionalist analysis of conscious experience.

The Alien Colour Effect

The data reviewed in the previous section tend strongly to the conclusion that word–colour synaesthesia is based upon an abnormal, probably genetically determined, projection hard-wired into the brain. Conversely, these data lend no support to the hypothesis that this condition results from any special form of associative learning. In the present section we report additional experimental data which further weaken the associative learning hypothesis. These data come from a study of a sub-group of word–colour synaesthetes who experience what we have termed the 'alien colour effect' (ACE; Gray, 1999). In this phenomenon the names of colours induce a colour experience that is different from the colour named. So, for example, the word 'red' might give rise to the experience of green, and so on. For a given word–colour synaesthete, the ACE may affect all, some, or no colour names.

As is the case for synaesthesia in general, the ACE appears to have been present for as long as the subjects can remember, that is, back to early childhood. Now, consider the opportunities for associative learning that this situation entails. A young child with the ACE would frequently encounter circumstances under which someone makes a statement of the form: 'see the red bus coming round the corner'. From statements such as these, the child has normal opportunities to learn the visual colour to which the word 'red' applies. Synaesthetes do indeed learn colour names normally: colour perception, as assessed by the Isihara colour plates, is normal, as is colour naming (Mattingley *et al*., 2001; Rich & Mattingley, 2002). We have not yet formally assessed these functions in our subjects with the ACE, but our informal observations suggest that they too are normal in these respects. Yet, in the example given above, as well as seeing a red bus come round the corner just after being told about the bus, the child with ACE would also experience a different colour, e.g. green, upon hearing the word 'red'. Thus she must frequently encounter opportunities for associative learning provided by chains such as: word 'red', experience of green, sight of red bus. If the first part of this chain, word 'red' followed by green experience, were due to associative learning in the first place, one would expect it to be unlearnt by these further associative learning opportunities. This, certainly, is what happens in countless experiments on counter-conditioning or reversal learning with both animal and human subjects. Thus, the existence of the ACE is incompatible with the associative learning account of word–colour synaesthesia.

Given the scope of the conclusions for functionalism that we seek to draw from the hard-wiring account of synaesthesia, we thought it important to validate the ACE experimentally. As in other recent reports of Stroop-like interference (see below) arising from synaesthetic experiences (for review, see Rich & Mattingley, 2002), we sought evidence that the ACE might delay the speed of colour naming. This experiment is reported for the first time here; we therefore present it in full.

Experiment 3: Colour Naming in Subjects with the Alien Colour Effect

We assessed a group of colour-word synaesthetes for the degree to which they displayed the ACE and divided them into groups accordingly. We calculated % ACE for each subject as the percentage of visually presented colour names which elicited 'alien' colour experiences. Words in which some letters elicited alien colours but others did not were weighted accordingly. We then measured their speed of colour naming in a conventional Stroop test. This compares the speed of naming the colour in which a row of 'X's is presented (as the control condition) to the speed of naming the ink colour in the 'Stroop' condition. In the latter, the name of a colour is presented (for example, the word 'red') that is incongruent with the colour of the ink (for example, green) in which it is written. The difference in speed of naming between the control and Stroop conditions provides a measure of the degree to which the colour name interferes with the processing of the ink colour. We also included a further condition in which the speed of colour naming is slowed even further, namely, 'negative priming' (e.g., Steel *et al.*, 2001). This is like the Stroop condition, but with the added complication that the correct colour response on trial *N* is the same as the one which has to be inhibited on trial *N-1*. We anticipated that, if the ACE as reported by the subjects is a real phenomenon, then in naming the ink colour they would suffer interference from the alien colour. This interference might in turn be exacerbated by the Stroop and/or negative priming effects. Thus, we predicted that, the greater the % ACE, the slower would be the naming of the ink colour, especially in the Stroop or negative priming conditions.

Design

A 3 x 4 mixed design was used. The single between-subjects factor was the 'Groups' variable with four levels: non-synaesthetes and 3 groups of synaesthetes with 0–35% ACE, 35–70% ACE and 70–100% ACE. The single within-subjects factor was 'Condition' (three levels: control, Stroop interference and negative priming). The control condition (CC) consisted of a row of coloured Xs; in the Stroop condition (SC) colour names were presented in ink colours which were (a) incongruent with the presented colour name and (b) different from the colour name on the previous trial; and in the negative priming condition (PC) colour names were presented in ink colours which were (a) incongruent with the presented colour name and (b) the same as the colour name on the previous trial.

Subjects

Forty-eight synaesthetes who reported seeing words in colour were selected from the database held by S. B-C. at the University of Cambridge. Forty-two were female and six were male. They had a mean age of 53.2, range 9–80, years. All subjects had responded to the Cambridge Synaesthesia Questionnaire, which comprised a list of 24 words, including: (a) 14 meaningful words in the semantic categories of objects and abstract terms; (b) 5 colour words — red, purple, brown, blue and green; (c) 5 first names, of people of both sexes. Subjects were required to write down the colour(s) they saw with each word. Responses to the colour words only were used to assess the alien-colour effect (ACE), as outlined above. Table 1 shows the resulting distribution of % ACE.

ACE	0% ACE	<50% ACE	50–99% ACE	100% ACE
No. of Ss	13	15	17	3

Table 1. The distribution of the Alien-Colour Effect (ACE) across all subjects

On the basis of these scores subjects were assigned to one of three groups: with 0–35%, 35–70% and 70–100% ACE. Ten subjects were tested from each group. The normal control group of ten subjects was selected from a sample of individuals from the Institute of Psychiatry and the general public. They all reported normal vision and no form of synaesthesia. The National Adult Reading Test (NART-revised; Nelson & Willison, 1991) was used to estimate verbal IQ, and the four groups were matched for this and age (Table 2; differences between groups non-significant by analysis of variance). The subjects were female with the exception of one male, and were all right-handed with the exception of two female subjects, one left-handed and one ambidextrous.

Group	Normal control (n=10)	0–35% ACE (n=10)	35–70% ACE (n=10)	70–100% ACE (n=10)
Age	42.7 (9.7)	55.20 (13.7)	53 (16.1)	56.20 (20.2)
PVIQ	119 (3.0)	121.6 (4.7)	121.1 (5.5)	120.5 (3.0)

Table 2. Age and Predicted Verbal IQ (PVIQ) of the four groups
(means and standard deviations in parentheses)

Materials

Stimuli (Steel *et al.*, 2001) were presented on an AST Ascentia portable computer. The subjects' reaction times (RTs) were recorded via a microphone and voice-key interfaced with the computer. The experimenter (S.C.) manually entered each vocal response made by the subjects using a four-button box, each button representing a different colour. The button-box was connected to the computer, which therefore recorded any errors made. Stimuli consisted of the words ‘BLUE’, ‘GREEN’, ‘YELLOW’ and ‘RED’, displayed in uppercase characters, and were displayed in blue, red, yellow or green ‘ink’.[4] Congruent pairings of names and colours (e.g., the word ‘GREEN’ in green ink) were excluded. A row of three, four, five, or six (randomly ordered) Xs was used for the control condition, again displayed in blue, yellow, green or red. All stimuli appeared in the centre of the screen, with subjects sitting approximately 75 cm from the screen. The pattern mask consisted of a grid containing the same four colours.

Procedure

Subjects responded to a sequence of 162 stimuli in total. The stimuli were ordered as 18 successive blocks, each consisting of 9 consecutive stimuli. Each block was allocated to one of the three experimental conditions (CC, SC or PC) in a quasi-random order fixed across subjects. Thus there were 6 blocks for each condition. Subjects were informed that coloured words and Xs would be displayed on the screen, and were instructed to name the ink colour in which the word or Xs were displayed as quickly as possible without making any vocal hesitations (these were recorded as errors by the experimenter). All subjects first carried out two practice trials, which were of the Stroop condition.

Each stimulus consisted of a black fixation cross displayed for 500 msec, followed by the target (i.e., the Xs or colour name). After 100 msec a pattern mask appeared and remained on display until the subject recorded a response via a voice-key. The next fixation cross was then presented immediately. This was repeated for the nine stimuli within each block. At the end of each block the subject’s mean RT and number of errors were displayed on the screen. The subject accessed the next block by pressing the space bar when ready. The first RT in each block was excluded in the analysis, as were RTs where an error had been made.

Results

The mean RTs for the three experimental conditions (control, Stroop and negative priming) are shown for each group in Figure 2 below. It can be seen that RTs for all the synaesthete groups were higher than for the non-synaesthetes. RTs were logarithmically transformed due to non-normality in the data set for the 70–100% ACE group in the CC condition. They were then submitted to a mixed 4 (Groups) x 3 (Conditions) analysis of variance, with polynomial contrasts on Groups ordered as in Figure 2 below. This revealed a significant overall difference between Groups ($F = 4.03$, d.f. = 3, 36, $p < 0.02$). A post hoc Bonferroni test showed a significant ($p = 0.008$) difference between the non-synaesthetes and the synaesthete group with 70–100% ACE. There was also a highly significant

[4] It would have been better in principle to run this experiment with colours matched to each synaesthete’s particular pattern of ‘alien colours’. However, this would have required a much more laborious experiment, which we therefore decided to run only if it turned out to be necessary to do so. In fact, as reported here, the standard Stroop test was good enough to demonstrate the interference effect we sought; and, indeed, as described, simply naming 4 Xs was sufficient. Thus the more laborious experiment (which would presumably demonstrate even clearer effects) was unnecessary.

(F = 11.29, d.f = 1, 36, p < 0.002) linear trend to slower RTs with increasing % ACE. Given that the mean ages of the groups (Table 1), although not significantly different, fell short of perfect matching, we re-ran this part of the analysis covarying for age. The critical linear trend remained significant, albeit with reduced significance (p = 0.02). The Condition effect was also highly significant (F = 198.79, d.f. = 2, 72, p < 0.001). Post-hoc t-tests showed that both the Stroop effect (difference between mean RTs in SC and CC conditions) and the negative priming effect (difference between PC and SC conditions) were significant (t = 16.11 and 4.0 respectively, df = 72, p < 0.001). There was no interaction between Group and Condition.

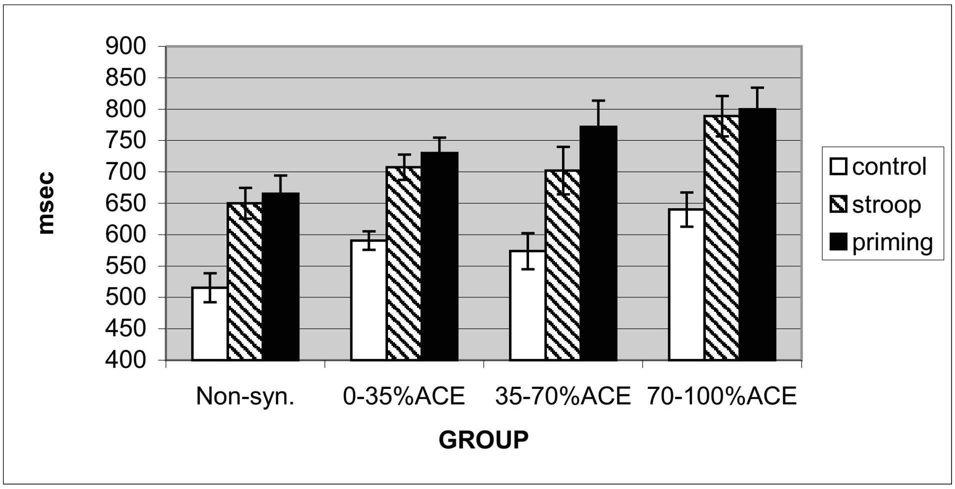

Figure 2. Mean reaction times and standard errors for the control, Stroop and negative priming conditions as a function of percentage ACE (alien colour effect) in Experiment 3. Non-syn: non-synaesthete group.

Discussion

These results confirm the reality of the self-reported ACE. The greater the percentage ACE reported, the slower was colour naming. This effect, furthermore, was observed as clearly in the control condition, in which the subject had only to name the 'ink' colour in which four Xs were presented, as in the Stroop and negative priming conditions. Note that the interference caused in colour naming by the ACE must precede the subject's overt utterance of the correct colour name. This interference is presumably due, therefore, to activation of the synaesthetic, incongruent colour by subvocal retrieval of the ink colour name. The additional conflict between ink colour and colour name inherent in the Stroop and negative priming effects was not required to bring out the effect of the ACE upon the speed of colour naming. Indeed, the Stroop and negative priming effects, as such, were unaffected by the ACE. Quantitatively, the degree of the ACE-induced slowing (if one compares full ACE colour naming speed to that of the non-synaesthetes; Figure 2) was about the same as the size of the Stroop effect itself.

The reality of the ACE, demonstrated in this experiment, casts further doubt on the possibility that word–colour synaesthesia could be the result of any associative learning process. Every time a colour name occurs in association with the perception of the colour named, and therefore also in conjunction with the alien colour experience triggered by the name, as presumably occurred in the experiment reported here, there is an opportunity for normal associative learning processes to reverse the aberrant association that putatively underlies the ACE. Yet the ACE persists unchanged from childhood to adulthood. It is extremely unlikely therefore that the ACE is established as the result of an initial stage of normal associative learning. By extension, it is also unlikely that word–colour synaesthesia in general rests upon such an associative basis.

Conclusions

Overall, the results of these experiments, together with the various strands of supporting data and argument above (and see also Ramachandran & Hubbard, 2001a), support the following conclusions.

(1) Word–colour synaesthesia does not result from aberrant associative learning.

(2) Word–colour synaesthesia is most likely due to an extra, abnormal, left-lateralized projection from cortical language systems to the colour-selective region (V4/V8) of the visual system.

(3) On this analysis, excitation in synaesthetes by heard or seen words of cortical language systems transmits excitation in an obligatory manner to the colour-selective region of the visual system.

(4) Activation of the colour-selective region of the visual system is sufficient to lead, automatically and involuntarily, to the conscious experience of colour, with the specifics of the colour experience depending upon the particular pattern of neuronal firing caused in V4/V8 by the excitation transmitted from cortical language areas.

(5) The occurrence of the synaesthetic colour experience in word–colour synaesthesia plays no functional role in relation either to speech or language perception or to colour vision. An intriguing gloss on this conclusion is provided by Ramachandran and Hubbard's (2001a, p. 26) description of a grapheme–colour synaesthete with anomalous colour vision who 'claimed to see numbers in colours that he could never see in the real world ("Martian colours")'. Such 'Martian' colours imply that, if a pattern of V4/V8 neuronal firing induced in synaesthesia differs from any elicited via the normal visual pathway, it can nonetheless give rise to a colour experience specific to the pattern *per se* and not to any visually linked functional relationships.

(6) The occurrence of the synaesthetic colour experience in the alien colour effect has behaviourally *dys*functional effects, as demonstrated in Experiment 3.

These conclusions, in our view, are incompatible with a functionalist account of word–colour synaesthesia. This condition provides a counter-example to what, at the outset of this paper, we called the 'complementary inference' from functionalism: namely, that, for any discriminable functional difference, there must be a corresponding discriminable difference between qualia. Within the behaviour of any given word–colour synaesthete there is a clear functional separation between the seeing of a colour presented via the normal visual channel, on the one hand, and the perception of that same colour triggered by a word. Yet, apparently, neither the qualia nor their neural bases (as tested in our fMRI experiments) produced by these two functional routes differ. Notice that these inferences are drawn on a *within-subject* basis. The same synaesthete subject experiences and reports on the qualia activated via the two different functional routes. This important feature of our experimental design eliminates the problem of the so-called 'privacy' of conscious experience. This problem would arise only if we were attempting to compare qualia between different subjects.

It is, of course, difficult to affirm a lack of difference in qualia with any certainty. It could be argued, for example, that the synaesthetic colour experience differs from normal colour experiences in at least two ways and therefore fails to provide a counter-instance to the complementary inference.

First, synaesthetic colour is always and necessarily combined with an auditory experience, that is, of the inducing word. Note, however, that these two experiences remain separate, as distinct from what happens in cases of perceptual 'binding', in which for example the shape, motion and colour of a moving object (a red kite, say, flying in the sky) are fully integrated into one visual percept even though each of these attributes is computed in a different brain region. The combination of word and colour in coloured hearing synaesthesia is more like the simultaneous experience of the sound of a violin and the sight of its being played. The two modalities of experience are tightly synchronised but perceptually separate. Thus there is no more reason to regard synaesthetic colour as not constituting a perceptual experience in its own right, separate from the associated auditory experience of the inducing word, than there is so to regard the sound and sight of the violin.

Second, the synaesthete is entirely aware of the different provenances of her colour experiences — surface colours by way of vision, synaesthetic colours by way of audition. But again there are parallels in non-synaesthetic experience. With careful selection of the appropriate shades, a red after-image can be induced by a green colour patch that is reported identical to a red patch directly presented. Yet the perceiving subject is in no doubt that one is an after-image and the other, a colour patch. From the latter fact it is not inferred that the subject is in error in reporting the two experienced colours to be identical. By the same token, if the synaesthete reports that a synaesthetic and surface colour are identical, the fact that she knows the different provenance of each should not count against the veridicality of that report.

Our synaesthete subjects do, in fact, report that their synaesthetic and 'real' experiences are closely alike. However, to examine this issue in greater detail, we have worked with a small number of word–colour synaesthetes with sufficient artistic talent to depict their colour experiences in response to specific words. We are currently applying fMRI to these subjects to determine just how closely the activation patterns elicited in V4/V8 by a given word and its corresponding picture resemble one another. This is a difficult experiment that may lie beyond the technical limitations of current neuroimaging techniques. But we hope that it will provide a route by which to test objectively this key assumption in our argument: that, in word–colour synaesthesia, similarity or even perhaps identity of qualia can occur despite disparate functional routes underlying them. Disproof of this complementary inference is not so damaging to functionalism as would be disproof of the primary inference: namely, that for any discriminable difference between qualia there must be an equivalent discriminable difference in function. But, we believe, it comes close.

There is an apparent escape hatch for functionalism in the finding that, in coloured hearing synaesthetes, left V4/V8 is devoted to synaesthetic colours and right V4/V8 to visually detected colours (Nunn *et al.*, 2002). The sensitivity of fMRI does not allow us to assert that this observation represents complete lateralized separation between the two functions. But, given that the different lateralizations were observed in the same subjects within a single scanning session (Nunn *et al.*, 2002), they cannot be dismissed as artefact. Thus one might try to salvage the functionalist account of coloured hearing synaesthesia by asserting that the two functions (elicited by spoken words or seen colours) do not in fact share qualia, since one is associated with qualia generated in left V4/V8 and the other, with qualia generated in right V4/V8. However, this line of defence must take as axiomatic what ought in our view to be an empirical hypothesis: namely, that different neural processing produces different qualia. Yet subjectively, to the synaesthete, both are experienced as colour. Indeed, one may also interpret the different lateralization of colour produced visually and synaesthetically as providing an even stronger counter-example to functionalism. For there is considerable evidence (Zeki, 1993) that activity in V4/V8 in either hemisphere is sufficient for the experience of colour. Thus an opponent of functionalism might argue that, in coloured hearing synaesthetes, colour experiences are produced by two routes which differ in *all* critical respects: input, output and the site (left or right hemisphere) of the strongest ‘neural correlate’ (Crick & Koch, 1995) of the consciousness of colour.

It may appear that, in adopting this joint line of attack upon functionalism, we are trying to have our cake and eat it too. The argument needs, therefore, to be spelt out carefully. There are three terms that have to be put together in any understanding of the relations between qualia (Q), functions (F) and brain processes (B). The complementary inference we have drawn from functionalism states that, if F1 differs from F2, then (provided that F1 and F2 belong to a domain of processing associated with qualia) F1 must be associated with Q1 and F2 with Q2, such that Q1 differs from Q2. As noted at the start of the article (p. 7 above), in most versions of functionalism functions are specified in terms of abstract processes alone (the box-and-arrow diagrams of cognitive psychology being a familiar example); however, in others they are specified in terms of actual neural processes in the brain. In the latter case, F1 is mediated by B1 and

F2 by B2. Assuming a case (like, so we claim, the case of word–colour synaesthesia) in which Q1 and Q2 are the same even though F1 and F2 differ, we can therefore envisage two possibilities: (1) that B1 and B2 do not differ, or (2) that they do. Both of these patterns of results run counter to the complementary inference and therefore to functionalism. However, they differ in that (1) places the fault line in functionalism between functions, on the one hand, and qualia-plus-brain processes on the other; whereas (2) places the fault line between functions-plus-brain processes, on the one hand, and qualia on the other. We anticipated the former outcome to our experiments. The latter, which is the result observed, is even more inimical to functionalism, in that qualia appear to be stripped by it of any necessary connection to *either* specific functions *or* specific brain processes.

Our findings also run counter to functionalist expectation in a second respect. Harnad (in press) has argued that qualia can be selected in biological evolution only in virtue of the fact that they are epiphenomenally linked to functions that have survival value. Synaesthesia calls for an evolutionary account, given the evidence that it runs in families and is likely to have a genetic basis (Ramachandran & Hubbard, 2001a; Harrison & Baron-Cohen, 1997; Marks, 1997). However, it is difficult to see, on Harnad's argument, how the ACE could ever arise. The understanding of language, in audition and vision, clearly has survival value, as does colour vision. One can also see that a neural linkage between language systems and colour vision could provide survival value, for example, by facilitating the naming of colours. But no natural account emerges along these lines of why this neural linkage should give rise to the perception of colours triggered by words in word–colour synaesthesia, an arrangement which is at best functionally neutral; and still less why it should give rise to the alien colour effect, which is (as we have shown) actively dysfunctional. Of course, this may only be a temporary state of affairs. It may simply be the case that the deleterious effects of the putative genetic mutation that gives rise to coloured have simply not yet had time to provoke strong negative selection. But it is difficult to see how any such deleterious effects can be exerted if there is no role for the conscious perception of the synaesthetic colours *in its own right.*

An anonymous referee of this paper has questioned the significance that this argument attaches to the fact that synaesthesia runs in families and is likely to have a genetic basis. As he points out, 'Huntington's chorea and breast cancer also run in families and are likely to have a genetic basis; but no-one would argue that they should be accounted for with an evolutionary account of their adaptive function. Many maladaptive traits and diseases have a genetic basis. Why should synaesthesia be any different?' There is indeed no reason why synaesthesia as such should not be deleterious, despite its being genetically determined (as the evidence suggests that it is). But the point we are making goes deeper than this. Functionalism supposes that qualia are fully dependent upon the functions with which they are associated. If that is so, it should not be possible for qualia *to compete negatively with those very same functions*. Yet, in the case of the alien colour effect, that is just what they appear to do.

There will perhaps be a temptation to dismiss our findings on the basis that they depend upon 'illusory' perception. We have ourselves, above, drawn a parallel between word–colour synaesthesia and other illusory experiences of colour and motion; in particular, they appear to rest upon the same neural foundation, namely, activation of that part of the visual system which is responsible for the analysis of the visual feature (colour, motion) concerned, without activation in earlier parts of the visual pathways. However, to dismiss our findings on this basis would be to misunderstand how normal vision works. In a very real sense, this too is illusory (Velmans, 2000). Thus, for example, in the particular case of concern to us here, that of colour vision, it is almost universally agreed that colours, as such, are not properties of the objects that we perceive as being coloured. The basis that such objects provide for the brain's construction of colours lies in the light reflectances of their surfaces as a function of the wavelengths of light that fall upon them (Zeki, 1993). There is no known relationship (other than correlational) between these reflectances, whether measured on the surfaces themselves or as computed by the brain, and the qualia by which they emerge into conscious perception. The phenomenon of word–colour synaesthesia provides an empirical basis upon which to ask an ancient philosophical question: why should not colour qualia have been used normally, as they are used unusually by word–colour synaesthetes, to represent in consciousness auditory inputs (words) rather than visual inputs (reflectances)? Perhaps coloured hearing synaesthetes are on the first step along an evolutionary pathway which could have led to the allocation of colour qualia to words (were it not for the fact that this pathway has been pre-empted by the visual system)?

This type of question, of course, takes us to the heart of the Hard Problem of conscious experience. Until we can go beyond correlation to mechanism in understanding how qualia come to be allocated to function, that problem will remain. The considerations advanced in this paper render it less likely that the allocation of qualia in word–colour synaesthesia is determined solely or even at all by function as such. And, given that functionalism purports to provide a completely general account of how conscious experiences relate to brain activity, even one such counter-instance, if it can be firmly established, may be sufficient to overthrow it.

Acknowledgements

We thank Craig Steel and Elke Haworth for permission to use their Stroop and negative priming tasks and Nanda Vythelingum for programming them. Our empirical research has been supported by the McDonnell Pew Foundation, and by a Consciousness Studies Award from the University of Arizona and the John E. Fetzer Institute. During the writing of this article JAG was a Fellow at the Center for Advanced Study in the Behavioral Sciences, Stanford University, with financial support from the John D. and Catherine T. Macarthur Foundation, Grant no. 32005-0. We are also grateful to Dominic ffytche, Dan Dennett and Max Velmans for helpful comments on the manuscript.

References

Aleksander, I (2000), *How to Build a Mind* (London: Weidenfeld & Nicholson).

Aleman, A., Rutten, G-J. M., Sitskoorn, M.M., Dautzenberg, G. and Ramsey, N.F. (2001), 'Activation of striate cortex in the absence of visual stimulation: An fMRI study of synesthesia', *NeuroReport*, **12**, pp. 2827–30.

Baron-Cohen, S., Harrison, J., Goldstein, L.H. and Wyke, M. (1993), 'Coloured speech perception: In synaesthesia what happens when modularity breaks down?', *Perception*, **22**, pp. 419–26.

Barnes, J. *et al*. (1999), 'The functional anatomy of the McCollough contingent colour aftereffect', *NeuroReport*, **10**, pp.195–9.

Bartels, A., Zeki, S. (2000), 'The architecture of the colour center in the human visual brain: new results and a review', *European Journal of Neuroscience*, **12**, pp. 172–93.

Borst, C.V. (ed 1970), *The Mind-Brain Identity Theory* (London: St Martin).

Bullmore, E.T. *et al.* (2001), 'Coloured noise and computational inference in neurophysiological (fMRI) time series analysis: Resampling methods in time and wavelet domains', *Human Brain Mapping*, **12**, pp. 61–78.

Buxton, R.B. (ed. 2002), *Introduction to Functional Magnetic Resonance Imaging: Principles and Techniques* (Cambridge: Cambridge University Press).

Crick, F. and Koch, C. (1995), 'Are we aware of neural activity in primary visual cortex?', *Nature*, **373**, pp. 121–3.

Dehaene, S., Kerszberg, M. and Changeux, J.P. (1998), 'A neuronal model of a global workspace in effortful cognitive tasks', *Proceedings of the National Academy of Sciences USA*, **95**, pp. 14529–34.

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

ffytche, D.H. (2002), 'Neural codes for conscious vision', *Trends in Cognitive Sciences*, **6**, pp. 493–5.

ffytche, D.H. *et al*. (1998), 'The anatomy of conscious vision: an fMRI study of visual hallucinations', *Nature Neuroscience*, **11**, pp. 738–42.

Gray, J.A. (1971), 'The mind-brain identity theory as a scientific hypothesis', *Philosophical Quarterly,* **21**, pp. 247–53.

Gray, J.A. (1987), 'The mind-brain identity theory as a scientific hypothesis: A second look', in *Mindwaves*, ed. C. Blakemore & S. Greenfield (Oxford: Blackwell).

Gray, J.A. (1999), 'The hard question of consciousness: Information processing versus hard wiring', in *Neuronal Bases and Psychological Aspects of Consciousness,* Vol 8., ed. C. Taddeo-Ferretto & C. Musio (Singapore: World Scientific).

Gray, J.A., Williams, S.C.R., Nunn, J. and Baron-Cohen, S. (1997), 'Possible implications of synaesthesia for the hard question of consciousness', in *Synaesthesia: Classic and Contemporary Readings*, ed. S. Baron-Cohen, J.E. Harrison (Oxford, Blackwell).

Grossenbacher, P.G. and Lovelace, C.T. (2001), 'Mechanisms of synaesthesia: Cognitive and physiological constraints', *Trends in Cognitive Sciences*, **5**, pp. 36–41.

Hadjikhani, N., Liu, A.K., Dale, A.M., Cavanagh, P. and Tootell, R.B.H. (1998), 'Retinotopy and colour sensitivity in human visual cortical area V8', *Nature Neuroscience*, **1**, pp. 235–41.

Hameroff, S.R. and Penrose, R. (1996), 'Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness, in *Toward a Science of Consciousness: The First Tucson Discussions and Debates*, ed. S.R. Hameroff, A.W. Kaszniak, A.C. Scott (Cambridge, MA: MIT Press).

Harnad, S. (in press), 'Turing indistinguishability and the blind watchmaker', in *Evolving Consciousness*, ed J. Fetzer, G. Mulhauser (London: John Benjamins).

Harrison, J.E. and Baron-Cohen, S. (ed. 1997), *Synaesthesia: A Review of Psychological Theories* (Oxford: Blackwell).

Howard, R.J. *et al*. (1998), 'The functional anatomy of imagining and perceiving colour', *NeuroReport*, **9**, pp. 1019–23.

Koch, C. and Crick, F. (2001), 'The zombie within', *Nature*, **411**, p. 893.

Marks, L.E. (1997), 'On coloured-hearing synaesthesia: Cross-modal translations of sensory dimensions', in *Synaesthesia: Classic and Contemporary Readings*, ed. S. Baron-Cohen, J.E. Harrison (Oxford, Blackwell).

Mattingley, J.B., Rich, A.N., Yelland, G. and Bradshaw, J.L. (2001), 'Unconscious priming eliminates automatic binding of colour and alphanumeric form in synaesthesia', *Nature*, **410**, pp. 580–2.

Moutoussis, K. and Zeki, S. (2002), 'The relationship between cortical activation and perception investigated with invisible stimuli', *Proceedings of the National Academy of Sciences, USA,* **99**, pp. 9527–32.

Nelson, H. and Willison, J.R. (1991), *National Adult Reading Test (NART): Test Manual* 2nd ed. (NFER-Nelson, Windsor, 1991).

Nunn, J.A., Gregory, L.J., Brammer, M., Williams, S.C.R., Parslow, D.M., Morgan, M.J., Morris, R.G., Bullmore, E.T., Baron-Cohen, S. and Gray, J.A. (2002), 'Functional magnetic resonance imaging of synesthesia: activation of V4/V8 by spoken words', *Nature Neuroscience*, **5**, 371–5.

O'Regan, J.K. and Noë, A. (2001), 'A sensorimotor account of vision and visual consciousness', *Behavioral and Brain Sciences*, **24**, pp. 939–73.

Paulesu, E. *et al*. (1995), 'The physiology of coloured-hearing: a PET activation study of colour-word synaesthesia', *Brain*, **118**, pp. 661–76.

Ramachandran, V.S. and Hubbard, E.M. (2001a), 'Synaesthesia: A window into perception, thought and language', *Journal of Consciousness Studies,* **8** (12), pp. 3–34.

Ramachandran, V.S. and Hubbard, E.M. (2001b), 'Psychophysical investigations into the neural basis of synaesthesia', *Proceedings of the Royal Society of London, B*, **268**, pp. 979–83.

Rich, A.N. and Mattingley, J.B. (2002), 'Anomalous perception in synaesthesia: A cognitive science perspective', *Nature Neuroscience Reviews*, **3**, pp. 43–52.

Searle, J.R. (1983) *Intentionality: An Essay in the Philosophy of Mind* (Cambridge: Cambridge University Press).

Searle, J.R. (1987), 'Minds and brains without programs', in *Mindwaves*, ed. C. Blakemore & S. Greenfield (Oxford: Blackwell).

Steel, C., Haworth, E.J., Peters, E., Hemsley, D.R., Sharma, T., Gray, J.A., Pickering, A., Gregory, L., Simmons, A., Bullmore, E.T. and Williams, S.C.R. (2001), 'Neuroimaging correlates of negative priming', *Brain Imaging*, **12**, pp. 1–6.

Tootell, R.B.H. *et al*. (1995), 'Visual motion after effect in human cortical area MT revealed by functional magnetic resonance imaging', *Nature*, **375**, pp. 139–41.

Velmans, M. (2000), *Understanding Consciousness* (London: Routledge).

Wager, A. (1999), 'The extra qualia problem: synaesthesia and representationism', *Philosophical Psychology,* **12**, pp. 264–81.

Zeki, S. (1993), *A Vision of the Brain* (Oxford: Blackwell Scientific Publications).

Zeki, S., Watson, J.D.G. and Frackowiak, R.S.J. (1993), 'Going beyond the information given: the relation of illusory visual motion to brain activity', *Proceedings of the Royal Society, London, B*, **252**, pp. 215–22.

Paper received July 2002