Anchoring Biases and the Cognitive Penetrability of Colour Experience

Raamy Majeed
University of Cambridge
DRAFT: 2017

ABSTRACT: Can cognitive states 'penetrate' our perceptual experiences? Macpherson (2012) argues that there is one alleged case of cognitive penetration that cannot be explained away, *viz.* the effects of cognition on colour perception, as demonstrated by Delk & Fillenbaum (1965). This paper aims to show that, though Macpherson's example is controversial, her arguments motivate a penetrability interpretation of several other experimental findings, especially if we understand these arguments as inferences to the best explanation. I demonstrate this by defending her argument from Zeimbekis's (2013) response, which claims that certain experimental factors, e.g. reduced acuity conditions, give rise to an anchoring bias that makes subjects misjudge (but not necessarily misperceive) colours. I argue that not only is the anchoring hypothesis compatible with the penetrability thesis, but that an explanation along the lines of anchoring that employs this thesis is a better explanation of the experimental results than one which doesn't. I thereby conclude that we still have abductive reasons to suppose that colour perception, in some cases, is affected by cognition.

1. Introduction

There are well-known experiments from cognitive psychology, which suggest that our perceptual experiences are cognitively penetrable: the phenomenal character of these experiences can be causally affected by our cognitive states.[1] There is also an ever-increasing set of responses, which seek to debunk any penetrability interpretation of these findings.[2] Most influential responses allude to methodological flaws of these experiments, e.g. they tend to only test hypothesis-confirming predictions, and thereby neglect hypothesis-disconfirming ones.[3] But since such flaws can be remedied, this type of debunking strategy leaves the question of whether there are genuine instances of cognitive penetration very much open, and a matter to be settled by future psychology.

There is, however, a strategy that reaches a more decisive verdict in favour of the penetrability sceptic. This is the judgement hypothesis: the relevant experimental data is

[1] This is to be kept separate from the thesis, discussed by Pylyshyn (1999), that cognitive states affect early visual processing. See Macpherson (2012) for an exposition of the difference.

[2] For a review of these experiments and responses, see Stokes (2013).

[3] E.g. see Machery (2015), and Firestone and Scholl (2015b).

explained as post-perceptual effects on judgement. Comparatively, this hypothesis is more damning than other debunking strategies, as it threatens to persist even if the methodological shortcomings of the experiments for cognitive penetration have been overcome. It is my view that there is an adequate way of responding to this hypothesis that has been overlooked. Subsequently, the aim of this paper is to clarify this response, and defend it from recent critiques.

The response I have in mind is due to Macpherson (2012), who argues that there is one alleged case of cognitive penetration that cannot be explained away, even by the judgement hypothesis, *viz.* the effects of cognition on colour experience, as demonstrated by Delk & Fillenbaum (1965). Though I think we have good reasons to suppose that this experiment too succumbs to methodological flaws (more on this later), her arguments are compelling, and it is instructive to re-assess the dialectic between her and her interlocutor, as a means to guard against the judgement hypothesis, wherever it may be invoked.

Since the purpose of this paper isn't to defend Macpherson's example, I won't be dealing with all the recent critiques of her proposal. Instead, I focus on the critique I find most compelling, namely the version of the judgement hypothesis proposed by Zeimbekis (2013). As we shall see, this proposal is compelling for two reasons. First, in Zeimbekis's hands, we get not just an assertion of the hypothesis, but an explanation for it as well. Second, Zeimbekis's proposal is directly aimed at addressing Macpherson's argument, and consequently focussing on the proposal helps expose both the strengths and weaknesses of her argument.[4]

There is one major piece of housekeeping we need to address before proceeding. The claim Macpherson originally wants to defend is that "there is one alleged case of cognitive penetration that cannot be explained away by the standard strategies one can typically use to explain away alleged cases" (2012: 24). Moreover, her opponents, including Zeimbekis, oblige by providing alternative explanations of this alleged case. The dialectic framed in this way is misleading, as it makes the penetrability thesis, i.e. that there are genuine instances of cognitive penetration, much too easy to defeat. For instance, there is no doubt that Zeimbekis succeeds in providing a plausible, rival explanation of the relevant empirical findings, which Macpherson thinks cannot be debunked. If this is all he has to do to settle the issue of cognitive penetration, the penetration sceptic has already won the debate.

But the proponent of the penetrability thesis need not concede defeat so easily. Given the empirical data underdetermines whether or not cognitive penetration occurs,

[4] Other responses to Macpherson include Gross *et al.* (2014), and Dokic and Martin (2015).

the dialectic is to be framed not in terms of whether alleged cases of cognitive penetration can be 'explained away' by alternative hypotheses, but whether the penetrability hypothesis or some rival is the best explanation amongst the set of explanations on offer. In other words, as Stokes (2015) assumes, the arguments for the penetrability thesis ought to be viewed as abductive arguments. Therefore, I shall proceed by treating Macpherson's argument as an abductive one, and the ensuing debate as a comparative one, concerned with finding the better explanatory hypothesis.

The paper is structured as follows. I outline Delk and Fillenbaum's experiment, and explain why Macpherson thinks we should prefer the penetration thesis to the judgement hypothesis with regards to explaining the experimental data (§2). I then sketch Zeimbekis's defence of the judgement hypothesis, and explain why both hypotheses, within the abductive context, are roughly on a par (§3). Finally, I offer a way we might break this stalemate, which favours the penetrability thesis (§4).

2. Explaining the Colour Perception Data

Delk and Fillenbaum (1965) conducted an experiment where subjects were asked to adjust the colour of a background patch to match the colour of two-dimensional cardboard cut-outs that were placed in the foreground. Some of these had the shape of characteristically red objects (a love-heart, a pair of lips, an apple), whereas others didn't (an oval, an eclipse, a mushroom etc.). All of these cut-outs were made from the same "orange-red" cardboard, whereas the colour of the background patch was controlled by twisting a knob, and could be made to go from shades of yellow to orange to red, or *vice versa*. Subjects were asked to tell the instructor to turn the knob to make it more red or yellow until the background colour matched that of the cut-out in the foreground. It was found that for cut-outs shaped as characteristically red objects, subjects selected a background that was redder than for those which weren't characteristically red. On this basis, Delk and Fillenbaum concluded that "past association of colour and form does in some way influence perceived colour" (pg. 293).

The experimental method used here is suspect on several grounds. For example, it doesn't adequately control for experimenter effects: having an experimenter control the knob is a classic case where the experimenter's expected outcomes might affect the results. More troubling is that Gross *et al.* (2014) failed to replicate these results when such methodological flaws were corrected and the experiment was run using virtual targets on computer displays instead of cardboard cut-outs. Since our aim isn't to vindicate the original results, but rather employ the experiment as a means of drawing

out the abductive arguments for and against cognitive penetration, we can set these worries aside.

Given our specific focus, the first thing to note is that, like all the other experimental findings that support the penetrability thesis, the data here underdetermines whether perceptual experiences are actually cognitively penetrated. To elaborate, Delk and Fillenbaum's conclusion goes beyond what their experiment actually shows: observers tend to assign typical colours to objects even when these objects do not have those colours. Call this the *empirical data*. This data is suggestive of an affect on perceptual experience, which is presumably why they claim that "the three red-associated figures was *seen* as redder than each of the other six figures" (pg. 293, my emphasis). Nevertheless, the data is also compatible with it being the case that the red-associated figures affect colour judgement without affecting colour perception itself. In other words, the data is compatible with the judgement hypothesis.

Since abductive arguments are comparative, let us proceed by seeing how this hypothesis stands with regards to the penetrability thesis. To begin, note that the debate is such that the point of comparison isn't between the penetrability and judgement hypotheses *per se*, but rather between the claims available to the proponents of these hypotheses with regards to explaining the empirical data. These claims give way to two further, and more specific, explanatory hypotheses. As I read her, Macpherson regards them as follows:

i) Subjects judge the colours of the (orange-red) foreground and modified (redder) background to be the same whilst experiencing them as different.

ii) Subjects judge the colours of the (orange-red) foreground and modified (redder) background to be the same because they experience them as the same.

Hypothesis (i) is the explanation of the results available to the proponent of the judgement hypothesis, and (ii) is that which is available to his opponent, the defender of the penetrability hypothesis. Both hypotheses attribute an error to subjects. But while (i) only attributes an error in colour judgement, (ii) also ascribes an error in perceptual experience. More specifically, the latter experiences involve 'inaccurate representations' of the relevant colours.

Macpherson raises two interrelated worries for (i), and *ergo* the judgement hypothesis. First, typical visual errors concerning misjudgement, e.g. in visual illusions, are due to perceptual inaccuracies where subjects' perceptual experiences misrepresent their surrounds. Such an explanation, however, isn't available to the interlocutor since he wants to deny a change in perceptual experience. Second, the error attributed to subjects

on the judgement hypothesis is actually greater than it initially appears, as it involves not just a misjudgement, but also a misjudgement in the face of accurate representation of colours by the perceptual experiences. These two worries taken together suggest that the explanatory demands of buying into the judgement hypothesis are too great for it to genuinely rival the penetrability hypothesis.

In actual fact, Macpherson goes further. She takes these two factors to show that buying into the judgement hypothesis involves attributing an error to subjects, which is "gross, brute and inexplicable" (pg. 42.). This is to overstate her case. Insofar as we grant that introspective judgements concerning the phenomenal characters of our experiences are fallible, we can explain (i) by appealing to such fallibility. This would, I grant, be an unlikely explanation for it attributes a systematic error in introspective judgement to all participants of the experiment. But an unlikely explanation is still an explanation, and so it isn't as if the proponent of the judgement hypothesis doesn't have any explanatory resources up his sleeve.

The strength of Macpherson's response rests on our taking it as an abductive argument. Subjects can, and sometimes, do err with regards to introspective judgements about what they perceive. Consequently, visual errors concerning misjudgement in the Delk and Fillenbaum experiment may be indicative of a misjudgement about perceptual experience, as opposed to a misperception itself. However, as Macpherson notes, in most cases, such visual errors are due to misperception. Thus, (ii) is still a more likely explanation of the empirical data than (i).

We should also not lose sight of Macpherson's second point: (i) attributes a gross error to subject in that they (allegedly) not only misjudge the colours, but they do so in the face of accurate perceptual representation of these colours. The proponent of the judgement hypothesis, thereby, needs a story as to why there is such an error, and it is doubtful that he can come up with one that rivals (ii) not just in its likelihood, but also its simplicity: subjects misjudge the colours simply because they misperceive them.

Macpherson, then, overstates her case, but not by much. The explanatory demands she raises for the proponent of (i) helps expose precisely why (ii) is a simpler and likelier, and therefore 'better', explanation of the empirical data than (i). So though the judgement hypothesis is a viable interpretation of the empirical data, we still have abductive reasons to favour the penetrability thesis.

3. The Judgement Hypothesis Reconsidered

The judgement hypothesis falters because the explanatory challenges to (i) are, arguably, too demanding to be sufficiently met. But what if (i) isn't the only way the proponent of

the judgement hypothesis can account for the empirical data? Zeimbekis argues that there is a second explanatory hypothesis available to its proponent, one that evades these explanatory challenges:

iii) Subjects judge the colours of the (orange-red) foreground and modified (redder) background to be the same (partly) because they fail to experience them as different.

Hypothesis (ii) claims that the colours of the foreground and modified background are experienced by subjects as being the same. It is possible to deny this, and yet hold that subjects (also) fail to experience these colours as being different.[5] Hypothesis (ii) and (iii) then come apart. Moreover, the way they do so proves significant, as it robs Macpherson of her critique of the judgement hypothesis. To elaborate, *contra* Macpherson, the proponent of the judgement hypothesis need not claim that subjects judge the colours to be different even though they look the same. He can make do by arguing that subjects simply fail to see the colours as being different, and thus make a judgement based on their prior knowledge concerning the stereotypical colours of the objects from which the cut-outs are shaped. This has the consequence of both sidestepping the explanatory demands, and preserving the cognitive *im*penetrability thesis.

But why believe (iii)? Hypothesis (iii) brings with it its own set of explanatory challenges, which Zeimbekis addresses head-on. He provides a two-pronged explanation, the first step explaining why the experimental conditions are such that subjects fail to discriminate the colour of the cut-outs from several possible background colours, and the second why subjects specifically choose the darker shades of red out of these background colours. It is these steps, taken in conjunction, which explain the empirical data.

3.1 Step 1

According to Zeimbekis, there are two reasons, which work 'cumulatively', to explain why subjects fail to experience the two colours as different. The first has to do with not the experiment itself, but how Delk and Fillenbaum interpret its findings. Let *F* be the colour of the cardboard from which the figures were cut, and *G* the mean value for the darker shades of red to which subjects matched the red-associated figures. Delk and

[5] According to Zeimbekis, "subjects are faced with successive background shades which are indiscriminable from one another, or with successive shades none of which can be discriminated with significant certainty from the color of the figure. There is no principled way to choose among these shades" (pg. 172).

Fillenbaum emphasis that the "mean differences between the red-associated figures and the other figures are not only highly significant statistically but also of substantial magnitude perceptually" (pgs. 292-293).

This is demonstrated by comparing *F* and *G* to regions of the Munsell colour space. Call the closest colours to *F* and *G* on the system *A* and *B*. (These matches were made by two independent judges). *A* is R/5/12 and *B* is R/4/12, where 'R' stands for hue, the first numerical value stands for saturation, and the second for chrome. The difference between *A* and *B* is one of saturation, where 4 is a darker shade of red than 5. Delk and Fillenbaum note that there is a considerable perceptual difference between *A* and *B*, by way of demonstrating a similar perceptual difference between *F* and *G*.

Zeimbekis's responds by challenging the above inference: "that *A* and *B* are discriminable does not imply that *F* and *G* are discriminable" (pg. 169). The wording here is unfortunate. Since *F* and *G* were mapped by independent observes to *A* and *B*, any perceptual differences between *A* and *B*, *ceteris paribus*, also *imply* differences between *F* and *G*. Nonetheless, this is easily remedied. Given that the colours in the Munsell system were only close approximations of the actual colours in the experiment, differences between the former don't *entail* differences between the latter. This claim about the lack of entailment suffices to make Zeimbekis's point, namely "*F* and *G* could turn out to differ phenomenally by much less than the two Munsell samples" (pg. 170).

Scepticism about there being a significant phenomenal difference between *F* and *G* gives us some motivation to believe (iii). Nevertheless, the clincher is this coupled with the second reason, which defeats the *ceteris paribus* clause. In the experiment, the cut-outs and the patch were dimly illuminated, and subjects had to look at them through an almost, though not entirely, transparent sheet of wax paper – the purpose of these being to blur the boundary edges between the shapes and their background. Macpherson takes these reduced acuity conditions to be an unimportant detail, whereas Zeimbeikis notes that they create precisely the kinds of conditions where observers could fail to discriminate colours, even ones that are phenomenally different under normal conditions.

These two reasons, then, act together to increase the likelihood that subjects fail to see any difference between the colours *F* and *G*. Moreover, according to Zeimbekis, this in turn provides adequate motivation for (iii).

3.2 Step 2

Hypothesis (iii) *qua* an explanation of the empirical data, however, is incomplete. The (supposed) fact that subjects fail to experience *F* and *G* as different can only partially

explain why subjects judge them to be the same. Here is why.

That subjects fail to see *F* and *G* as different is compatible with subjects failing to see the difference between *F* and several other colours, some lighter than *F*, and some equivalent to it. In fact, if the reduced acuity conditions are efficacious, we would expect there to be a (vague) colour boundary, stretching from ones lighter than *F* to ones darker, whose colours are all phenomenally indistinguishable from *F*. So we need an explanation for why subjects specifically choose *G* out of all the possible colours within this colour boundary. The aim of step (2) is to provide such an explanation.

According to Zeimbekis, subjects choose *G* because the experimental conditions create an anchoring effect: "in borderline cases about which no principled judgement can be made … observers tend to resolve uncertainty by using bias which favours initial values" (pg. 172).

The basic idea of anchoring, which originates from Tversky and Kahneman (1974), is that when subjects make judgements in conditions of uncertainty, they use a heuristic, which biases them in favour of the starting point (or anchor). For example, say I ask, 'how many islands are there in the Maldives?' and give one set of subjects the initial value 55, and another 750. The answers to this question would be indicative of an anchoring bias provided subjects were genuinely uncertain as to the number of islands in the Maldives, and each group estimated a median value closer to their initial one.

For Tversky and Kahneman, anchoring works as a form of insufficient-adjustment. In borderline cases where subjects are provided with an anchor, subjects make insufficient adjustments such that their judgements are anchored by the initial value. Zeimbekis exploits this notion of anchoring as insufficient-adjustment to (partially) explain the empirical data.

The experiment conditions were such that subjects were made to make a judgement in a borderline case; the figures, for instance, were 'orange-red'. Moreover, in some instances, they were provided with an initial value of 'red' in the form of cut-outs that were of stereotypically red objects. It was then found that, while subjects sometimes physically over-adjusted the colour of the background patch (from yellow to a very dark red), they didn't sufficiently deviate from the starting point. Subsequently, the experimental results can (partially) be explained as an anchoring effect.

I will have more to say about anchoring biases in the next section. For now, note that the anchoring explanation, which makes up step (2), isn't presented as a way of responding to Macpherson's explanatory challenge. This is taken care of by step (1). Rather, step (2) aims to explain why subjects choose the darker shade of red, out of a set of possible colours they could have chosen as a match. These steps, as I understand Zeimbekis, are individually necessary, and jointly sufficient, to explain the empirical data.

In the scheme of understanding Macpherson's argument as an abductive one, it is still uncertain whether hers or Zeimbekis's hypothesis is the best explanation of the empirical data. Zeimbekis himself provides no comparative analysis, and it is far from obvious that one is a better explanation than the other. Nevertheless, the strength of Zeimbekis's response is that it appears to provide us with a genuine, and plausible, rival to the penetrability hypothesis. Which has the consequence of robbing the penetrability theorist of her example, and with it, any decisive reason to favour the penetrability thesis.

4. The Penetrability Hypothesis Reconsidered

It appears, then, that we have reached a stalemate between the proponent of the penetrability hypothesis and her rival. In this section, I attempt to break the stalemate, and in a way that (tentatively) favours the penetrability thesis. I do so by making the following two criticisms of Zeimbekis's judgement hypothesis.

First, I challenge the assumption made in step (2) that anchoring concerning colour judgement is a genuine rival explanation of the empirical data to that of cognitive penetration. Second, I challenge the features in step (1), which makes Zembekis's version of the judgement hypothesis immune from Macpherson's explanatory demands. These criticisms are independent, but together, they suggest that the penetrability thesis still stands on better explanatory footing than the judgement hypothesis.

4.1 Anchoring as a Form of Cognitive Penetration

Prima facie, the notion of anchoring is a very different hypothesis to the penetrability thesis. This difference, however, belies the fact that they might be distinct descriptions of two inter-connected phenomena. Why? Because the anchoring effect with regards to colour judgement doesn't rule out cognitive penetration. If anything, the effect is best explained with reference to it. Let me explain.

The role cognitive penetration is supposed to play is delegated to a feature of the anchoring mechanism at the initial stages of the experiment, when the anchor is first introduced:

> [W]hen subjects perceive a heart shape of the same color, a recognitional concept with a memory-color is triggered. This anchors judgements of the figure's color in the concept red: the color will be more likely to be classified and thought of as red, although it could just as well have been classified as orange. (Zeimbekis 2013: 172)

Note that this explanation of the anchoring effect is compatible with cognitive penetration. Both the interlocutors can grant that a recognitional concept is triggered. In

fact, the proponent of the penetrability thesis needs some cognitive state, e.g. a recognitional concept, to do the penetrating, so it's hardly as if she would deny that such a concept is triggered. Hence, to suppose that the concept doesn't affect experience is to beg the question against the proponent of cognitive penetration.

Zeimbekis doesn't beg the question in this way. Instead, his claim is that we don't require cognitive penetration as anchoring suffices as an explanation of the empirical data. My charge against him isn't that there is no anchoring effect here. Anchoring, as Tversky and Kahneman note, is amongst the most robust and easily replicable findings in psychology. *Prima facie*, I see no reason why the Delk and Fillenbaum results can't be described as an anchoring effect. The problem is in its implementation as an explanation of these results. As Strack (1992) points out, anchoring itself is a descriptive notion as opposed to an explanatory one, and thereby can't be employed as an explanatory concept unless we understand the psychological mechanisms that underlie it. My charge against Zeimbekis consists of two claims. First, the mechanisms typically invoked to explain anchoring don't explain the kind of anchoring that is supposed to occur in the colour experiment. Second, cognitive penetration is one way to explain this particular anchoring effect. These claims prove significant as their conjunction illustrates that an explanation of the empirical data which employs an anchoring effect grounded in cognitive penetration is a better explanation than one that merely appeals to an (unexplained) anchoring effect. Consider each of these claims in turn.

Tversky and Kahneman themselves, as we saw, propose an anchoring mechanism involving insufficient-adjustment. Most of the contemporary psychological literature on anchoring, however, is sceptical of this explanation. One reason this is the case is that, as Mussweiler and Strack (2001) argue, it is hard to explain why adjustment occurs when it comes to plausible anchors because they provide no reason to make adjustments in the first place. Thus, it appears that the scope of the insufficient-adjustment explanation seems limited to anchors that are implausible, which is problematic because anchoring effects are obtained whether or not the anchor values are implausible. To this we can add that the colour experiment, in particular, is a hard case for the anchoring as insufficient-adjustment notion to explain, as it appears to be a case where the anchor is plausible: presumably there is nothing implausible about giving a red-associated figure the anchor value 'red'.

The reason for being doubtful that other standard mechanisms for anchoring can explain the empirical data also has to do with the anchoring value of the experiment, but in combination with how this value is introduced. Most, if not all, cases of anchoring mentioned in psychology involve anchors that are numerical values. This is why some explanations of the phenomenon include mechanisms such as numerical priming and

conversational inferences. By contrast, the anchor value in the colour experiment isn't numerical. Rather it appears to be something perceptual itself, or at least something inherently tied to perception. There is nothing in the psychological literature, which adequately accounts for these kinds of anchors.

The standard explanation, and the last of the anchoring mechanisms on offer, is selective accessibility.[6] The basic idea being that comparing the judgemental target to the anchor value increases the accessibility of the anchor-consistent subset of target knowledge. For example, suppose subjects were provided with the anchor value 55 for the question, 'how many islands are there in the Maldives?' According to the selective accessibility account, these subjects would be more likely to test the possibility that the target is actually 55. (They may do so by selectively retrieving knowledge from memory that is consistent with this value, e.g. 'It's a small nation'). Subsequently, anchor-consistent knowledge about the Maldives is increased, which in turn biases the final answer in a way that favours the anchor.

Again, this is proposed as an explanation for anchors that are numerical. Nonetheless, we can see how such a mechanism might explain the (alleged) anchoring effects of the colour experiment. Subjects that select the anchor value 'red' are more likely to consider potential targets that are consistent with this anchor. Which in turn increases the likelihood that they will select a background shade consistent with this anchor.

This explanation, however, won't suffice to explain the empirical data. Typical anchoring paradigms concern anchors that are explicitly provided instead of self-generated. This is why Mussweiler and Strack, in proposing their selective accessibility account of anchoring, state that we need an additional explanation in these other paradigms as to why subjects self-generate the anchors that they do. Mussweiler and Strack themselves suggest that self-generation of anchors is explained by the mechanisms typically invoked to explain anchoring itself. That is, anchors might be selected based on conversational inferences, numeric priming or insufficient adjustment.

Nonetheless, these mechanisms seem as ill equipped to explain the self-generation of the anchor 'red', as they are to explain the anchoring effects of the colour experiment. Consider conversational inferences. As part of their experiment, Delk and Fillenbaum tested whether instructions biased colour-matching judgements. There were, in actual fact, three groups of subjects: Group (1) were given both colour and figure names (e.g. "the yellowish-orange oval", "the reddish apple"), Group (2) were given just the figure names, and Group (3) were given neither. They found no effects of instructions on

[6] See Mussweiler, Englich and Strack (2004).

colour judgement, which in turn suggests that conversational inferences weren't required to self-generate the anchor 'red'.

Cognitive penetration provides one way to fill this explanatory void. Subjects self-generate the relevant anchor because seeing the red-associated figures trigger a recognitional concept with the memory-colour red, and this in turn has an affect on how they perceive the colours of these figures: they are seen as redder than they actually are. It is for this reason that colour-matching at the later stages of the experiment is anchored by 'red'.

Of course, it is also possible that cognitive penetration over-determines the selection of the anchor, as the triggering of the concept might suffice — i.e. along with the experimental conditions, which are relevant for the anchoring effects to take place. But if a sort of conceptual priming is all that is required, we would expect some judgemental differences on the basis of the differing instructions, as mentioning the colours would also trigger the colour concept, and in a way that makes it more readily available. For instance, we would expect the anchoring effects to be greater for those primed with the colour labels than just the colour-associated shapes. A lack of such effects, however, suggest that what's doing the work isn't conceptual priming *per se* but the cognitive penetration of perceptual experience itself.

More work is needed to adequately establish whether cognitive penetration plays a role in anchoring colour judgements. But by the same token, more work is also needed to establish whether anchoring itself actually plays a role in colour judgement. For now, we can conclude that, as things stand, an explanation of the anchoring effects that relies on cognitive penetration is a better explanation of the empirical data than one that doesn't.

4.2 The Causal In-Efficacy of the Masking Effects

We have just seen that step (2), the step that is supposed to (partly) explain the empirical results, might actually involve cognitive penetration. Now we turn to step (1), which makes Zembekis's version of the judgement hypothesis immune from Macpherson's explanatory demands. The hypothesis rests on the plausibility of (iii): subjects judge the colours of the (orange-red) foreground and modified (redder) background to be the same (partly) because they fail to experience them as different. Here I challenge this, and with it the plausibility of the judgement hypothesis.

To recap, Zeimbekis offers us two reasons that work cumulatively to explain why subjects fail to experience the two colours as different. First, he exploits the fact that the colours in the experiment are only approximates of those of the Munsell colour system

to show that whilst there might be phenomenal differences between the relevant colours in the system, this doesn't entail that there is any such differences between the colour of the red-associated figures and the redder background shades to which they are matched. Second, he argues that certain features peculiar to the Delk and Fillenbaum experiment, *viz.* the reduced acuity conditions, create precisely the kinds of conditions where observers could fail to discriminate colours, even ones that are phenomenally different under normal conditions.

The first reason illustrates why it is possible that subjects could fail to see the colours of the figures and those of the background as different, but it is the second that lends this plausibility. Any credibility we give to (iii), thereby, is reliant on the second reason. The trouble is, this reason won't help explain away alleged cases of cognitive penetration where no masking effects are present.

To elaborate, as we saw, there are independent methodological problems with the Delk and Fillenbaum experiment, which should make us treat their findings with caution. So the strengths of the judgement and penetrability hypotheses, as both interlocutors grant, rests not on their ability to explain the colour experiment findings *per se*, but on their ability to explain a whole host of empirical results that are indicative of cognitive penetration.[7] The problem is that a lot of these findings result from experiments without reduced acuity conditions, or any similar conditions that mask the relevant phenomenal differences between the colours compared.

Consider the familiar series of experiments undertaken by Levin and Banaji (2006). In these trials, subjects were shown faces with stereotypically black and white features, but in the same grey-scale, with the same surface luminance. Moreover, they were asked to match these faces to a grey patch, which could be altered (by the subjects themselves) from darker to lighter shades. It was found that stereotypically white faces were matched to lighter shades of grey than those that were stereotypically black. On this basis, it was concluded that, "relatively abstract expectations about relative reflectance of objects can affect their perceived lightness" (pg. 501).

Penetrability sceptics do offer ways of debunking the experimental results, e.g. Firestone and Scholl (2015a) exploit differences in the distribution of surface luminance between the two groups of faces to explain the experimental results *sans* any top-down effects by cognition. It is my view that these strategies falter because they fail to account for the variety of results confirmed, e.g. similar results were obtained when a racially ambiguous face was labeled either 'black' or 'white'. This isn't the place to explore this

[7] E.g. they both attempt to explain the results from Hansen *et al.* (2006), which show that the task of adjusting coloured figures to grey on computer displays differ for colour-neutral and colour-associated figures.

disagreement. In any case, it is beside the point. What is telling is that there were no reduced acuity conditions in these experiments, nor any analogous conditions, which could have masked the colour differences between the faces and the patch. Penetrability sceptics certainly offer no examples of masking effects as part of their debunking strategies. Hence (iii), though still possible, is a less likely explanatory hypothesis than Zeimbekis makes out.

Perhaps I am being unfair, as what Zeimbekis thinks is essentially doing the explanatory work isn't necessarily the reduced acuity conditions, or even any masking effects more generally, but the fact that subjects in these alleged cases of cognitive penetration "are placed in situations of judgmental uncertainty" (pg. 174). For he continues:

> More compelling grounds for accepting the cognitive penetrability of color perception would be if subjects were (or could be) placed in conditions which preserved both conceptual priming and adequate conditions for pairwise matching, and again performed differently for color-associated and for neutral objects. (Zeimbekis 2013: 174)

The charge, then, must be that whilst alleged cases of cognitive penetration, to which the judgement hypothesis is relevant, preserve conceptual priming, they aren't placed in adequate conditions for subjects to make the relevant matches. It is this reason, which lends (iii) credibility.

Whether this demand itself is fair depends on what we take to be the 'adequate conditions' for pairwise matching. If these conditions are ones where masking effects are absent, we see that there are several alleged cases of cognitive penetration that satisfy this demand. If, however, they are conditions where there is no ambiguity, no judgemental uncertainty, whatsoever, this seems to be too high a demand. For it might be the case that perceptual experiences are only, or most commonly, penetrated by cognition precisely when there is judgemental uncertainty.

Nothing about the penetrability thesis itself, nor the assumptions made by their proponents, as the experiments they point to illustrate, suggests that instances of cognitive penetration need be free of any judgemental uncertainty whatsoever. Cognitive penetration, whether rare or ubiquitous, is said to occur in our day-to-day lives. Given that it is very rarely that conditions of our lives preserve ideal conditions for pairwise matching, we should not expect experiments that seek to test the penetrability thesis to preserve such conditions either.

Several experiments that imply cognitive penetration, then, meet adequate conditions (by reasonable standards) for pairwise-matching.[8] This in turn suggests that their findings are not a result of subjects failing to experience the colours, which require pairwise matching, as different. There are two lessons to be drawn from this. First, insofar as masking effects play no role in other alleged cases of cognitive penetration, including ones to do with pairwise colour-matching tasks, we have reason to doubt that they do so in the colour experiment. Hence, the explanatory hypothesis (iii), though possible, is found to be implausible. Second, even if the masking effects of the experiment were causally efficacious in the way that Zeimbekis claims, we see that (iii) looses its broad appeal. That is, something analogous to (iii) can't be invoked to explain other alleged cases of cognitive penetration, even ones where the judgement hypothesis might (initially) prove to be a suitable explanation.

The significance of these lessons should not be underestimated. If (iii), and ones like it, are found to be implausible, the defender of the judgement hypothesis would once again be at a loss to meet Macpherson's explanatory demands. He would, again, have to grant that subjects make judgements that aren't reflected in, and indeed go against, their perceptual experiences. Thus, in terms of the comparative analyses, which abductive arguments call for, we see that the penetrability thesis still comes out as a better explanation of the empirical results.

[8] These include ones often cited in the philosophy of perception literature, e.g. Levin and Banaji (2006), but also ones less familiar, e.g. see the experiments mentioned in Athanasopoulos *et al.* (2009), especially Thierry *et al.* (2009).

References

Athanasopoulos, P., Wiggett, A., Dering, B., Kuipers, J.R., and Thierry, G. (2009). 'The Whorfian Mind: Electrophysiological Evidence that Language Shapes Perception'. *Communicative & Integrative Biology* 2(4): 332-34.

Delk, J. L. & Fillenbaum, S. (1965). 'Differences in Perceived Color as a Function of Characteristic Color'. *The American Journal of Psychology* 78: 290–3.

Firestone, C. & Scholl, B. J. (2015a). 'Can you experience 'top-down' effects on perception?: The case of race categories and perceived lightness'. *Psychonomic Bulletin & Review* 22: 694-700.

----- (2015b). 'Cognition does not affect perception: Evaluating the evidence for 'top-down' effects'. *Behavioral and Brain Sciences*.

Gross, S., Chaisilprungraung, T., Kaplan, E., Menendez, J. A., & Flombaum, J. (2014). 'Problems for the Purported Cognitive Penetration of Perceptual Color Experience and Macpherson's Proposed Mechanism'. *The Baltic International Yearbook of Cognition, Logic and Communication* 9: 1-30.

Hansen, T., Olkkonen, M., Walter, S. and Gefenfurtner, K. R. (2006). 'Memory Modulates Color Appearance'. *Nature Neuroscience* 9 (11): 1376-8.

Levin, D. T., & Banaji, M. R. (2006). 'Distortions in the Perceived Lightness of Faces: The Role of Race Categories'. *Journal of Experimental Psychology: General* 135(4): 501-12.

Machery, E. (2015). 'Cognitive Penetrability: A No-Progress Report'. In J. Zeimbekis & A. Raftopoulos (ibid): 57-74.

Macpherson, F. (2012). 'Cognitive Penetration of Colour Experience: Rethinking the Issue in Light of an Indirect Mechanism.' *Philosophy and Phenomenological Research* 84: 24-62.

Mussweiler, T. & Strack, F. (2001). 'The Semantics of Anchoring'. *Organizational Behaviour and Human Decision Processes* 86: 234-255.

Mussweiler, T., Englich, B. & Strack, F. (2004). 'Anchoring Effects'. In R. Pohl (ed.), *Cognitive Illusions: A handbook of fallacies and biases in thinking, judgement and memory*. London, UK: Psychology Press: 183-200.

Pylyshyn, Z. (1999). 'Is Vision Continuous with Cognition?: The Case for Cognitive Impenetrability of Visual Perception'. *Behavioral and Brain Sciences* 22: 341-365.

Stokes, D. (2013). 'Cognitive Penetrability of Perception'. *Philosophy Compass* 8(7): 646-63.

---- (2015). 'Towards a Consequentialist Understanding of Cognitive Penetration'. In J. Zeimbekis & A. Raftopoulos (ibid): 75-100.

Strack, F. (1992). 'The different routes to social judgements: Experiential versus informational strategies'. In L. Martin & A. Tesser (eds.), *The Construction of Social Judgements*. Lawrence Erlbaum: 249–275.

Thierry, G., Athanasopoulos, P., Wiggett, A., Dering, B., and Kuipers, J.R. (2009).

'Unconscious Effects of Language-specific Terminology on Pre-attentive Color Perception'. *Proceedings of the Natural Academy of Sciences* 106: 4567-70.

Tversky, A., & Kahneman, D. (1974). 'Judgment under uncertainty: Heuristics and biases'. *Science* 185, 1124–1131.

Zeimbekis, J. (2013). 'Color and cognitive penetrability'. *Philosophical Studies* 165: 167-75.