Forthcoming in *Philosophical Studies*

DOI: 10.1007/s11098-016-0784-2

The final version is available at <http://link.springer.com/article/10.1007/s11098-016-0784-2>

**Attention to Mental Paint and Change Detection[[1]](#footnote-1)†**

Assaf Weksler

Department of History, Philosophy, and Judaic Studies

The Open University of Israel

[assaf.weksler@gmail.com](mailto:assaf.weksler@gmail.com)

Abstract: *According to the influential thesis of attentional transparency (AT), in having or reflecting on an ordinary visual experience, we can attend only outwards, to qualities the experience represents, never to intrinsic qualities of the experience itself, i.e., to “mental paint.” According to the competing view, attentional semitransparency (AS), although we usually attend outwards, to qualities the experience represents, we can also (perhaps with effort) attend inwards, to mental paint. So far, philosophers have debated this topic in strictly armchair means, especially phenomenological reflection. My aim in this paper is to show how to design an experiment, using the change detection paradigm for studying visual working memory that, if yielding positive results, would support AS. The structure of the argument is as follows. In standard change detection tasks, which involve attention to qualities the experience represents, there exists a known object integration effect (Luck and Vogel 1997). I formulate a hypothesis I call “the turn-off hypothesis,” according to which this effect does not exist in change detection tasks where subjects are instructed to attend to mental paint. I argue that, given AS, we have some reason to expect the turn-off hypothesis to be true, but given AT, we do not. Consequently, the truth of the turn-off hypothesis would support AS. I further argue that one can straightforwardly test the turn-off hypothesis, taking the standard change detection paradigm and instructing participants to attend inwards.*

Keywords: mental paint, qualia, attention, transparency, attention to mental paint, change detection, psychology of attention, neuroscience of attention, intentionalism, visual working memory

1. **Introduction**

According to the influential thesis of attentional transparency (AT; often the thesis is simply called “transparency”), in having or reflecting on an ordinary visual experience, we can attend *only* to qualities the experience attributes to external objects, such as colors and shapes. In other words, we can attend only to qualities the experience represents, or to qualities figuring in the (representational) contentof the experience (Harman 1990; Tye 2002). For convenience, I will sometimes call these qualities “external qualities.” On this thesis, we cannot attend to “mental paint” (sometimes called “qualia”). Mental paint is supposed to be an intrinsic quality of experience that constitutes (at least part of) its phenomenal character, i.e., it constitutes “what it is like” to have the experience (see Section 2).

According to attentional semitransparency (AS; this is my term), while we *usually* attend to external qualities, we *can* direct our attention inwards, to mental paint (Burge 2003; Kind 2003, 2008; Loar 2003).[[2]](#footnote-2) Some proponents of AS (Kind and Loar), though not all (Burge), think that attending to mental paint requires a special method. Block (2003) appears to accept AS[[3]](#footnote-3), but this is not entirely clear because of his suggestion that we can be aware of mental paint without attention (see Section 2). Papineau (2014) appears to accept AS, but not for cases of illusion and hallucination (see Footnote 9 below).

The debate between AT and AS is central to contemporary philosophy of mind (see Section 2). So far, however, philosophers have conducted it via strictly armchair considerations, specifically phenomenological ones. It appears that philosophers have neglected to ask whether and how vision science can shed some light on this attentional debate.

My aim in this paper is to show how to design an experiment, using the change detection paradigm for studying visual working memory that, if yielding positive results, would support AS.

The basic argument of the paper is as follows. In standard change detection tasks, which involve attention to qualities the experience represents, there exists a known object integration effect (Luck and Vogel 1997). I formulate a hypothesis I call “the turn-off hypothesis,” according to which this effect does not exist in cases where subjects are instructed to attend to mental paint. I argue that, given AS, we have some reason to expect the turn-off hypothesis to be true, but given AT, we do not. Consequently, the truth of the turn-off hypothesis would support AS. I further argue that the turn-off hypothesis can be empirically tested, by taking the standard change detection paradigm and adding the manipulation of instructing the participants to attend inwards.

Note that while the truth of the turn-off hypothesis would *support* AS, the falsity of the turn-off hypothesis would not count *against* AS (at least not straightforwardly).[[4]](#footnote-4) Let me explain. That the turn-off hypothesis is false means that the object integration effect remains, when participants try to attend inwards. Doesn’t this count against AS? After all, AS gives us reason to expect the object integration effect to disappear in such a case, yet it doesn’t. The answer is that this does not straightforwardly count against AS. The problem is that there are a lot of alternative ways – consistent with AS – of explaining why the object integration effect remains. I discuss some of them in sections 4 and 5 below.

The plan of the paper is as follows. In Section 2, I provide some background about AS and mental paint. In Section 3, I describe the turn-off hypothesis and the experiment for testing it, arguing that positive results would support AS. In sections 4 and 5, I address objections.

1. **Background: Mental paint and attentional semitransparency**

The AT/AS debate informs a central debate in the metaphysics of experience, between intentionalists (or representationalists) and supporters of the mental paint view (sometimes called “qualia realists”). There are several ways to spell out what intentionalism amounts to, some stronger than others, but the differences between them are irrelevant to my concerns. I will describe the strongest version, which is also the simplest. According to it, the phenomenal character of sensory experience is identical to the intentional (representational) content of the experience. The content, moreover, is “Russellian”, i.e., it consists of properties the experience attributes to external objects, and not of modes of presentationsof such properties (Mendelovici 2014; for an influential weaker version of intentionalism, see Tye 1995). For example, suppose I look at a red tomato. My experience represents the tomato as red; it attributes to the tomato the property of redness. According to the present version of intentionalism, this representational fact *exhausts* the way the experience of redness feels to me. There is nothing in the phenomenal character of experience aside from its attributing to the tomato the property of redness. Specifically, one need not add mental paint or a mode of presentation of redness or anything else to capture fully what it is like to have the experience.

Supporters of the mental paint view claim that intentional content does *not* exhaust the phenomenal character of experience since *intrinsic properties* of the experience also contribute to the phenomenal character. Philosophers call intrinsic properties of the experience that contribute to the phenomenal character “qualia” or “mental paint[[5]](#footnote-5).” In short, the mental paint view says that mental paint exists (Block 2003; Burge 2003; Kind 2003, 2008; Loar 2003; Papineau 2014; Prinz 2012).

Let me describe one plausible way to spell this out within a physicalistic framework[[6]](#footnote-6). An experience of red is roughly a population of cortical neurons firing in a certain way. This firing of the population of neurons has representational content: it represents redness (via some causal connection to red things in the world). The firing pattern itself counts as a “vehicle” of the content. On the mental paint view, this vehicle is red mental paint (Papineau 2014)[[7]](#footnote-7): it is responsible for the way the experience feels to us, at least in part. On the intentionalist view, in contrast, this vehicle is irrelevant to the way the experience feels to us; only the content is relevant.

For example, according to Prinz (2012, ch. 4), the vehicles of color experiences are populations of neurons in V4 firing in a certain temporal pattern he calls a “waveform” (he later complicates the picture, claiming the vehicles are “vectorwaves,” which are constellations of waveforms). On Prinz’s account, this firing pattern is not sufficient to generate a *conscious* color experience. He claims that In order for the experience to be conscious, it must be attended[[8]](#footnote-8) (see Section 5 below). Whether or not attention is necessary for consciousness is orthogonal to my concerns. So I don’t mind accepting a view on which these waveforms in V4 are the vehicles of conscious color experiences even when unattended (which appears to be, roughly, Lamme’s 2003 view). Combining this with the mental paint view yields the result that red mental paint is a certain (attended, if Prinz is right) waveform in V4.

When I argue below that AS gives us some reason to expect the turn-off hypothesis to be true, I assume in the background roughly the preceding account of vehicles of color experience. While I think it is possible to extend the analysis to cover other accounts of these vehicles, such as accounts focusing on functional role (cf. Papineau 2014) or accounts focusing on “global broadcast” (see Section 5 below), doing so will take me too far afield.

So far, in this section, I have described the mental paint view and intentionalism. I will shortly connect this debate to the issues of attentional semitransparency (AS) and attentional transparency (AT). Before doing so, I would like to quote Burge (2003), as he describes AS in a vivid way:

“Of course our attention is normally focused on the redness of the tomato when we are seeing a tomato. That is because physical objects and their properties are the primary objects of our perceptual systems. […This] does nothing to show that someone who can conceptually distinguish the tomato's redness from the characteristic quality of his perceptual *experience* of the tomato's redness [i.e., from red mental paint] is unable, with appropriate prompting, to *attend* to the latter. It seems to me that doing so is no great feat.” (pp. 405-6, emphases in the original)

Some philosophers appear to hold that the mental paint view implies AS (equivalently, AT appears to imply intentionalism). Here is why. Mental paint is a phenomenal aspect of our experience. So if it exists, we should arguably be able to attend to it. As Loar (2003, p. 80) puts it, “I do not know what to make of the idea of a phenomenal quality that cannot be directly attended to.” Thus, the mental paint view appears to imply straightforwardly that (1) we can attend to mental paint. Now, on the mental paint view, (2) our experience does not represent mental paint[[9]](#footnote-9). Moreover, (3) usually, we attend to qualities our experience represents, and not to mental paint[[10]](#footnote-10). From (1)-(3), it follows that AS is true. Thus, the mental paint view apparently implies AS. Burge (2003), Kind (2003, 2008), Loar (2003) and Papineau (2014; but see Footnote 9 above) appear to share this line of thought[[11]](#footnote-11).

Block (2003) disagrees. He considers seriously the possibility that mental paint exists yet we cannot attend to it. On this possibility, we can be aware of mental paint without attending to it. He thinks this possibility makes sense because he holds that in general conscious awareness does not require attention. If he is right, then the mental paint view does not imply AS. I will not try to assess Block’s proposal here. The fact that many leading friends of mental paint appear to hold that the existence of mental paint implies AS is enough for making the study of AS important. Incidentally, Block’s settled view appears to be that, at least sometimes, we *can* attend to mental paint, in ordinary experiences (see Footnote 2 above). It thus appears that Block accepts AS, despite denying that it follows from the mental paint view.

Before closing the section, let me describe one of the methods recommended by philosophers for attending to mental paint (remember that if Burge is correct, no special method is needed). Loar (2003) devotes a full-length paper to such a method. I will describe his basic idea concisely. Suppose you are having an experience of a red tomato, and you are attending to the redness your experience represents. Here is the recipe for attending to red mental paint. First, imagine a situation in which (a) you are having an experience E with the same phenomenal character as the one you are now having, and (b) E represents greenness, in the sense of being systematically caused by green objects (as in Block’s 1990 inverted earth story). Second, attend to what your current experience and the imagined experience have in common. The attended common factor, according to Loar, is red mental paint. So when you are attending to the common factor, you are in fact attending to red mental paint, according to him.

1. **The turn-off hypothesis and how to test it**

**3.1 The turn-off hypothesis**

I now describe Luck and Vogel’s (1997) influential change detection experiments, which will lead us to formulate the turn-off hypothesis. Luck and Vogel’s experiments are specially designed to test the capacity of visual working memory. In a paradigmatic experiment, participants viewed a sample array of (two to six) objects with four features each (gap, size, orientation, and color) as depicted in Figure 1, followed by a brief delay (a blank screen), and a target array. In simple feature tasks, the subjects were instructed to indicate whether a specific feature type, such as size, has changed while ignoring all the other features. In conjunction tasks, the subjects were instructed to detect a change *in either feature*.

|  |
| --- |
|  |
| *Figure 1.* A sample array and a graph of results representative of Luck and Vogel’s (1997) classic change detection experiments. The graph shows a comparison of performance in different tasks, as a function of the number of objects in the display (set size). The different tasks are to detect a gap change, a size change, an orientation change, a color change, or a change in either feature (a conjunction task). The performance is nearly identical in call cases. Source: Luck and Vogel (1997), with permission from Nature Publishing Group. |

Results: the change detection accuracy was nearly identical in both kinds of task: very high up to four objects, and dropping quickly afterward (see Figure 1). This result was surprising, because when viewing four objects, a simple feature task involves remembering only four features (one feature per object), but a conjunction task involves remembering sixteen features (four features per object). Hence, the nearly identical performance in both kinds of tasks means that “16 features distributed across 4 objects can be retained as accurately as 4 features distributed across 4 objects” (Luck and Vogel 1997, p. 280). Call this “the object integration effect.”

Luck and Vogel, as well as other researchers, explain the object integration effect in terms of something like an “object file” held in visual working memory (Treisman 2006; Luck and Vogel call it “an integrated object percept”). Each file is a sort of an object-specific short-term memory, containing representations of external features of a single object. In visual working memory, there is room for roughly four object files (Luck and Vogel 1997). Creating the files and maintaining them in working memory requires one to attend to the features in question, which are features the experience represents, throughout the experiment, including the blank screen stage (Treisman 2006; Johnson et al. 2008)[[12]](#footnote-12). This requirement is an instance of the general principle that maintaining representations of features in working memory requires attending to the features (Carruthers 2015). Let me emphasize that while feature-integration into some object representation *can occur* *without attention*, the resulting object-representation does not enter working memory (hence, it does not count as an object file), meaning it does not persist through the blank-screen delay and is consequently unhelpful in change detection tasks (Treisman 2006).

That visual working memory involves object files, each storing representations of features of external objects, makes sense, since features in the world are in fact organized into objects, and it is crucial for the organism to track and recognize these objects.

Suppose you are looking at a red square, attending to the redness and squareness your experience represents. As I have explained, it makes sense for the visual system to integrate representations of redness and squareness into an object file, since redness and squareness are in fact (or are probably) properties of a single object. Now suppose, with AS, that you can also attend to mental paint, i.e., to the *vehicles* of these two representations, namely to two types of neuronal firing patterns, in different cortical areas, as described in Section 2. Unlike the properties they represent, these two vehicles are *not* properties of a single object, as they are simply different neuronal firing patterns, located in different brain areas. Thus, creating an object file with representations of these vehicles involves an error, since the file represents the two vehicles as features of the same object, which is wrong. Consequently, it does not make sense for the visual system to integrate representations of vehicles (of experiences) systematically into object files, whenever one attends to mental paint, as doing so involves a systematic creation of an erroneous representation. In other words, a systematic error made by the visual system is an anomaly; hence, it is unlikely: it is not something we expect. Thus, absent a story that predicts the systematic error in question, we have reason to expect the visual system not to integrate representations of vehicles of experiences into object files (in Section 5, I discuss a story that predicts the systematic error). Conclusion: prima facie, we have reason to expect that attending to mental paint (if such a form of attention exists) will not lead to a creation of an object file that stores representations of it. Consequently, the object integration effect should disappear. In short, we have reason to expect that attending to mental paint (if such a form of attention exists) will “turn off” the object integration effect.

If we *couldn’t* attend to the redness and squareness the experience represents, then it would perhaps make sense for the visual system to create an object file containing representations of the vehicles of these representations, as a substitute for the real thing, despite the systematic error involved. It would be a sort of a useful error. However, according to AS, we usually attend to the redness and squareness the experience represents. Hence, there is no need to create an object file representing the vehicles in question. Creating such a file is therefore not even a *useful* error.

Note that I am not saying that we have *overwhelming* reason to expect that attending to mental paint will turn off the object integration effect. Perhaps the visual system, in fact, makes the aforementioned systematic error for some reason (see Section 5). We cannot at present, prior to experimenting, be sure that it does not. I only claim that we have enough reason to motivate us to test the matter experimentally.

Consider now the following empirical hypothesis (I focus on Loar’s method for concreteness. My analysis carries over straightforwardly to other methods).

*The turn-off hypothesis.* In change detection tasks, when participants follow the recipe for attention to mental paint suggested by Loar, there is no object integration effect.

In light of the above, given AS, and given that Loar’s method for attending to mental paint works, we have some reason to expect that the turn-off hypothesis will come out true. For, if Loar’s recipe works then participants who follow it end up attending to mental paint, and we have some reason to expect that attention to mental paint will turn off the object integration effect.

Consequently, if the turn-off hypothesis comes out true, AS is in a good position to explain why. The explanation will be basically as follows. Because it is erroneous to create object files representing vehicles of experience, the visual system stores representations of such vehicles *discretely* in visual working memory, detached from object files, and this is why the object integration effect is missing when subjects are instructed to follow Loar’s recipe.

In contrast, AT seems to give us no reason to expect the turn-off hypothesis to be true. For, if AT is true, then Loar’s method does not lead to attending to mental paint (no method does!). So, despite following Loar’s method, the participants are still attending to qualities the experience represents. Thus, we expect to find an object integration effect, as always.

The upshot is that experimentally confirming the turn-off hypothesis would support AS over AT because AS can provide a good explanation for the truth of the turn-off hypothesis, whereas AT apparently cannot.

**3.2 A sketch of an experiment for testing the turn-off hypothesis**

How do we test the turn-off hypothesis? The answer seems straightforward: we use the familiar Luck and Vogel paradigm except we instruct the participants to follow Loar’s recipe for attending to mental paint, at the beginning of each test. We want to see whether this manipulation turns off the object integration effect. If it does, the turn-off hypothesis is thereby confirmed.

An illustration: the participants look at two kinds of simple features, say color and orientation, spread over two to six objects. We first instruct them to attend, using Loar’s method, to all the color mental paint in their experience. When they are confident that they have succeeded, they press a button, making the screen go blank for a short period, after which a target array appears. The participants are instructed to indicate whether anything has changed, color-wise. We do the same for orientation[[13]](#footnote-13), and for the conjunction of color and orientation. In the conjunction case, the participants are required to attend to all the color mental paint and all the orientation mental paint in their experience at the same time.

We then compare the performance in simple feature cases (colors only or orientation only) with performance in the conjunction case. What we are looking for is the absence of an object integration effect. In other words, we expect that performance will be sensitive only to the *number of features* required to remember, regardless of whether they are arranged in conjunctions (i.e., two features of one object) or are isolated (i.e., one feature per object). For example, suppose performance in the simple feature cases is reliable up to four objects, i.e., four features in total (one feature per object). In that case, we expect performance in the conjunction case to be reliable up to two objects, since each of them has two features, meaning that the participants need to remember four features in total. Finding this would confirm the turn-off hypothesis.

It might be that Loar’s method, which involves imagining hypothetical situations, will distract participants, thereby disrupting change detection performance to some extent, regardless of whether or not they successfully attend to mental paint. I would like to clarify that such a diminished performance does *not* amount to a confirmation of the turn-off hypothesis. For, the turn-off hypothesis does not say that when subjects follow Loar’s method, performance is *diminished*. The hypothesis says that when subjects follow Loar’s method, *the object integration effect* disappears, which is something else entirely.

To repeat, to confirm the turn-off hypothesis, we need to find that when participants follow Loar’s method, performance becomes *insensitive* *to whether features are organized into objects* on the screen or not. Consequently, it does not matter whether Loar’s method makes subjects reliably detect fewer features in comparison with ordinary change detection tests (in which participants don’t use Loar’s method). What matters is whether Loar’s method makes the number of reliably detected features in the simple feature tasks *match* the number of reliably detected features in the conjunction task.

A worry: Loar’s method is a complicated one to do, and it appears I require participants to perform it multiple times simultaneously. Specifically, in conjunction tasks involving four objects with four features each, participants apparently need to apply it *sixteen times simultaneously*. This procedure looks overly difficult. We cannot expect participants to do it.

To address the worry, I describe an application of a version of Loar’s method to multiple objects and features, which is not overly difficult (for participants who grasp the inverted earth idea well). This version of the method differs in detail, but not in essence, from the version I have described in Section 2. I use a first person description to make it easy for the reader to imagine herself performing the steps I describe.

I am facing the conjunction task of remembering both orientation mental paint and color mental paint. I look at the display in Figure 1. I understand Loar’s idea that my experience could feel the same yet represent different features, on inverted earth (note that I do not have to accept the idea, only understand it). Reflecting on my experience, it is clear to me which qualities of my experience will remain constant while representational content shifts, according to Loar. I judge (with understanding), “*These color qualities* and *these orientation qualities* are supposed to remain constant, while representational content shifts.”[[14]](#footnote-14) Arguably, knowing which qualities a demonstrative (such as “these color qualities”) picks out involves attending to them (Campbell 2002). If so, then when forming the judgment in question, which has demonstratives in its content, I inevitably attend to the qualities they pick out (which are mental paint, according to Loar). Thus, forming this judgment amounts to applying Loar’s method to the entire display at once, instead of to each feature individually. After forming the judgment, I press a button, making the screen go blank for about a second. Then a new display appears, and I need to indicate whether or not anything has changed.[[15]](#footnote-15)

A complication: the capacity of working memory determines not only the number of features one can remember but also the number of features one can think about, or form a judgment about, at a given moment (Carruthers 2015). Suppose (without loss of generality) the capacity of working memory for mental paint is four. This implies that when looking at the display in Figure 1, I can form a judgmentabout four mental paint tokens at most[[16]](#footnote-16), at a given time. Thus, if Loar’s method works, then when I judge, “these color qualities and these orientation qualities are supposed to remain constant…,” the demonstratives do not pick out *all* the twelve mental paint tokens I experience (six colors and six orientations); they pick out only four. Consequently, applying Loar’s method in the way I have suggested results in me attending to only four mental paint tokens out of twelve, at a given time.

This is not a problem, however. There is no need for me to attend to more than four mental paint tokens at a given time, because (ex hypothesi) I can *remember* at most four. Attending to more than four mental paint tokens will have no effect on change detection performance because that performance depends on the maximum number of features I can remember, not on the maximum number of features I can attend to.

In sum, in this section I have argued that the turn-off hypothesis, if true, would support AS, because the latter gives us some reason to expect the former to be true, whereas AT (apparently) does not. Further, I have sketched an experiment for testing the turn-off hypothesis, manipulating the change detection paradigm.

1. **An objection from divided attention**

I have claimed that we have reason to expect that if one directs attention to mental paint, no object file will be formed (because forming it is erroneous). Consequently, there will be no object integration effect. One might present the following objection: perhaps instructing subjects to attend to mental paint will result in *divided attention*, i.e., attention both to mental paint and to external qualities. If enough attention is directed to individual external qualities, and not merely to the external scene as a whole (cf. Treisman 2006), an object file for external qualities will be formed. Consequently, there will be an object integration effect.[[17]](#footnote-17)

Response: the idea of divided attention is not a threat to the main thesis of this paper, which is that *if* the experiment I have suggested confirms the turn-off hypothesis it will support AS. While it’s true that as far as we know now, trying to attend inwards might lead to divided attention, which, in turn, might lead to the creation of an object file, we are not sure that this will happen. We have *some* reason to expect such results, but at the end of the day, what we need to do is simply carry out the experiment I have suggested and see what happens.

If we conduct the experiment and see that the object integration effect is gone, we will have confirmed the turn-off hypothesis. Such a result would support AS. For, as I have explained above, AS can explain the truth of the turn-off hypothesis whereas AT (apparently) cannot.

If we conduct the experiment and see that the object integration effect is still in place, we will have *dis*confirmed the turn-off hypothesis (or at least failed to confirm it). The story about divided attention will then provide a possible explanation for the falsity of the turn-off hypothesis, which is consistent with AS. This explanation could help block the inference from the falsity of the turn-off hypothesis to the falsity of AS.

It thus seems that the story about divided attention leaves the main thesis of this paper unharmed.

But doesn’t the divided attention story weaken the *motivation* for conducting the experiment in the first place? Not in a significant way. True, *if* we had very strongreason to think that attending inwards will lead to divided attention, and that the spilled-over attention to external properties will be large enough and focused enough to generate object files, then we would be very confident that the turn-off hypothesis is false without experimenting. If this were the case, carrying out the experiment would be redundant. Such a consequence would be fatal for the project of this paper.

Fortunately, we *don’t* have very strong reason to think that attending inwards will lead to divided attention and that the spilled-over attention will be significant enough to generate object files. We only know that the divided attention story *might* be true, i.e., that it is not out of the question. This is not enough to render the experiment redundant.

1. **Objection: the redirection hypothesis**

I have argued that, given AS, and given that Loar’s method works, one has some reason to expect the turn-off hypothesis to be true because one has some reason to expect attention to mental paint to turn off the object integration effect. The reason is that an object integration effect with respect to mental paint requires the visual system to create objects files storing representations of it, i.e., storing representations of the *vehicles* of experience. However, such object files involve a systematic error, as they represent various vehicles of experience as features of a single object, which is wrong. Moreover, such a systematic error is unlikely to occur.

This argument presupposes that the visual system does not make systematic errors, except in special circumstances. In other words, the default position is to expect the visual system to avoid systematic errors. We need a special reason to move away from this default position.

An objector might claim that we *have* such a special reason. It seems plausible that attention to mental paint is a matter of the familiar mechanism of attention to the content of experience being redirected to mental paint, i.e., to the vehicle of experience. Call this “the redirection hypothesis.” We know that the familiar mechanism of attention to experiential content, i.e., to qualities the experience represents, makes representations of these qualities enter object files. So, plausibly, when one redirects this mechanism to the vehicles of experience, it leads to similar results, namely to representations *of the vehicles* entering object files. The idea is that the mechanism of attention always works in the same way: it always integrates representations of the attended “thing” into object files. Thus, despite involving a systematic error, creating an object file representing vehicles of experience is a *natural* *behavior* *we should expect* if the redirection hypothesis is true.[[18]](#footnote-18)

In short: the objector grants that, absent a special reason to think otherwise, we should expect the visual system not to integrate representations of vehicles into object files. The objector then adds that we *have* a special reason to think otherwise. The reason is that the redirection hypothesis is plausible, and it implies integration of representations of vehicles into object files. Thus, all things considered, we do not have a reason to expect the object integration effect to shut down, when one attends to mental paint, or so the objector argues.

Response: the objection is dialectically similar to the divided attention objection. The objector presents a plausible possibility (the redirection hypothesis) that plausibly implies that when one attends inwards, the object integration effect persists. However, since the possibility is merely *plausible*, not something we are certainabout, it merely tells us that we have *some* reason to expect the object integration effect to persist. It is not enough to undermine our reason to expect the object integration effect to disappear. The upshot is that we simply have two reasons pulling in opposite directions. We have some reason to expect the object integration effect to disappear (the systematic error consideration) and some reason to think it will persist (the redirection hypothesis). This does not undermine our motivation to carry out the experiment I have suggested. On the contrary: it *invites* theexperiment, as it could help settle the dispute.

Analogous to the case of divided attention, if the experiment *disconfirms* the turn-off hypothesis, the redirection hypothesis will block inference from this outcome to the *falsity* of AS. But if the experiment *confirms* the turn-off hypothesis, it will *support* AS.

Perhaps an objector would claim that the redirection hypothesis is not only plausible but certain (given AS), and that the implication from the redirection hypothesis to the existence of an object integration effect, when one attends inwards, is certain as well. After all, isn’t it clear that there is only one system for visual attention, and that it always works in the same way, leading to feature integration? If this is right, the objector continues, then we know in advance that the turn-off hypothesis is false, making my suggested experiment *redundant*, which (as emphasized before) is fatal for the project of this paper.

To defuse this objection, I will provide a consideration against the redirection hypothesis, relying on ideas from the neuroscience of attention and consciousness. The consideration is not meant to be conclusive: my aim is only to cast doubt upon the redirection hypothesis, thereby showing that it is far from certain. This doubt shows that we cannot be sure, at present, that the redirection hypothesis is true and, consequently, we cannot be sure, before carrying out the experiment, that the turn-off hypothesis is false.

Consider the following quite general and widespread thesis from the neuroscience of attention. The neural mechanism of attention to a property one’s visual experience represents is an amplification of the firing of the population of neurons representing that property, in visual areas of the brain (Carruthers 2015; Fries et al. 2001; Fries 2005; Prinz 2012). On one influential version of this theory, the amplification consists in the firing of the population of neurons becoming synchronized in the gamma band (“gamma synchrony”), i.e., 35 to 90 hertz (Fries et al. 2001; Kuznetsova & Deth 2008; Prinz 2012). The synchronization allows the content of the firing neurons to reach other brain areas (see Fries 2005 for a specific hypothesis about how this works). The important point is that attention to any property one’s experience represents indeed involves the same mechanism, on this approach. Attention to a square involves neurons representing squareness firing in gamma synchrony. Attention to redness involves neurons representing redness firing in gamma synchrony. Attention to a house involves neurons representing properties of the house firing in gamma synchrony, and so on.

I have briefly spelled out the neuroscience of visual attention to qualities the experience represents. I would like to consider what this implies for the (prospective) neuroscience of attention to red mental paint (if such a form of attention exists). To do so, I once again take on board the view of color experiences along the line of Prinz (or Lamme), mentioned in Section 2. On this view, an experience of redness is a firing population of neurons in V4 (e.g., a waveform), that represents redness, and that meets certain further conditions (e.g., is attended).

Suppose now that R1 is a population of neurons in V4, which fires in a certain pattern, and which consequently represents redness. The specific firing pattern of R1 is the *vehicle* of this visual representation of redness. Call this vehicle VEH1. Suppose that R1’s firing in the VEH1 pattern meets whatever conditions Lamme or Prinz require for it to become phenomenally conscious. Consequently, VEH1[[19]](#footnote-19) *is* red mental paint. Given the aforementioned theory of attention, the neural mechanism of attention to the redness experience represents (i.e., to the *content* of R1, when firing in VEH1), is gamma synchrony of R1 (or, more precisely, gamma synchrony of its firing pattern VEH1).

Now consider: what is the neural mechanism of attention to red mental paint, namely to VEH1? It can’t be a matter of R1 (and VEH1) being synchronized in the gamma band because this is the neural mechanism of attention to the redness experience represents. This strongly suggests that *something else* involving R1 (and VEH1) occurs when we attend to red mental paint (if we can attend to mental paint). Instead of gamma synchrony, perhaps attending to red mental paint involves R1 synchronously firing in lower frequencies (e.g., the alpha band, or 8-12 hertz)? Perhaps it is not a matter of synchrony at all, but some other form of amplification of R1? The answer does not matter for present purposes (although it is an interesting avenue of research in its own right). What matters is that, apparently, attention to mental paint (if such a form of attention exists) involves *a different neural* *mechanism* than attention to qualities the experience represents.

On its face, this casts doubt upon the redirection hypothesis. That the neural mechanisms of the two kinds of attention are different suggests that attending to mental paint is *not* a case of taking the good old mechanism of attention to the content of experience and *redirecting it* to mental paint, i.e., to the vehicle of experience.

In arguing that the neural basis of attention to mental paint is something other than gamma synchrony, I have assumed Prinz’s or Lamme’s theory of phenomenal consciousness. But the argument is more general. It applies to *any* theory of phenomenal consciousness on which the vehicle of conscious red experiences (i.e., red mental paint, if there is such a thing) is some firing pattern of a population of neurons in visual cortex, representing redness. Without loss of generality, let me call this pattern "P." The argument goes through no matter which conditions P must meet in order to be conscious. For, the heart of the argument is that P's firing in gamma synchrony cannot underlie attention to P itself (i.e., to red mental paint) because neuroscience tells us that P’s firing in gamma synchrony underlies attention to the *content* of P.

I have explained that the argument generalizes to any theory of consciousness on which the vehicle of conscious red experiences is P. But which theories other than Prinz’s and Lamme’s fit this description? Apparently, the global workspace theory of consciousness (Baars 1988; Dehaene & Naccache 2001) does. Let me elaborate.

Prinz and Lamme hold that visual representations can be conscious without being stored in working memory. Proponents of the global workspace theory of consciousness disagree. On that theory, a conscious representation simply *is* a representation that is stored in working memory, and is consequently available to various distant brain systems (hence is “globally” available) via long-range “workspace neurons.”

Suppose one looks at a red object, and consequently P, which represents redness, is activated. Suppose further that this representation *gets stored in working memory*. In neuroscientific terms, this means that prefrontal and parietal brain areas maintain P in an active and amplified state, for a relatively long period, independently of external stimulus (see the recent review by Gazzaley and Nobre 2012). According to Dehaene and Naccache (2001), being in this amplified state for a sufficient duration activates workspace neurons, which make the content of P globally available.

The global workspace theory suggests, therefore, that the vehicle of conscious red experiences is P, once it meets a certain condition. The condition is that prefrontal and parietal areas maintain it in the way described above. I have claimed that my argument showing that gamma synchrony does not underlie attention to mental paint applies to any theory on which the vehicle of conscious red experience is P, no matter which conditions P must meet in order to be conscious. It thus appears that the argument applies to the global workspace theory.

1. **Conclusion**

In this paper, I have formulated “the turn-off hypothesis”, according to which, in change detection experiments, when participants are instructed to attend to mental paint using Loar’s method, there will be no object integration effect. I have argued that (a) the turn-off hypothesis is empirically testable via an experiment I have sketched and that (b) its truth would support AS over AT. I have emphasized that the situation is notsymmetric: while confirming the turn-off hypothesis would support AS, disconfirming the turn-off hypothesis would *not* refute AS (at least not straightforwardly), because there are many ways consistent with AS of explaining why the turn-off hypothesis is false.

The basic argument for (b) was that an object integration effect for mental paint requires creating object files in visual working memory that represent it, i.e., that represent *vehicles* of experience. Object files of this sort systematically represent different vehicles as features of a single object, which is wrong. It is unlikely that the visual system would perform such a systematic error. Consequently, it is reasonable to expect, in light of AS (but not in light of AT), the object integration effect to disappear when one follows the instruction for attending to mental paint.

I have considered the objection that if attending to mental paint results in divided attention, the object integration effect will persist, because of the attention outward. I have explained that we do not know in advance whether attending to mental paint yields divided attention. We should carry out the experiment and see whether the object integration effect disappears when participants follow the instruction for attending to mental paint. If it does (i.e., if the experiment confirms the turn-off hypothesis), we will get support for AS. If it does not, the divided attention story is one way of blocking the inference from this experimental result to the falsity of AS.

I have also considered the objection that creating object files for vehicles of experience is something the visual system will likely do *if* it is the upshot of taking the familiar mechanism for attention to the content of experience and redirecting it to the vehicle of experiences (the redirection hypothesis). In response, I have provided neuroscientific considerations that cast doubt upon the redirection hypothesis. The upshot is that we do not at present know whether the redirection hypothesis is true. Consequently, as with the divided attention case, we should carry out the experiment and see whether it confirms the turn-off hypothesis. If it does, we will get support for AS. If it does not, the redirection hypothesis is one way of blocking the inference from the falsity of the turn-off hypothesis (or from the failure of confirming it) to the falsity of AS.

**Bibliography**

Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge, Mass.: Cambridge University Press.

Block, N. (1990). Inverted Earth. *Philosophical Perspectives*, *4*, 53–79.

Block, N. (2003). Mental Paint. In M. Hahn & B. Ramberg (Eds.), *Reflections and Replies: Essays on the Philosophy of Tyler Burge*. Cambridge, Mass.: MIT Press.

Burge, T. (2003). Qualia and Intentional Content: Reply to Block. In M. Hahn & B. Ramberg (Eds.), *Reflections and Replies: Essays on the Philosophy of Tyler Burge*. Cambridge, Mass.: MIT Press.

Campbell, J. (2002). *Reference and Consciousness*. Oxford: Oxford University Press.

Carruthers, P. (2015). *The Centered Mind: What the Science of Working Memory Shows Us about the Nature of Human Thought*. Oxford: Oxford University Press.

Dehaene, S., & Naccache, L. (2001). Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework. *Cognition*, *79*(1-2), 1–37.

Fries, P., Reynolds, J. H., Rorie, A. E., & Desimone, R. (2001). Modulation of Oscillatory Neuronal Synchronization by Selective Visual Attention. *Science (New York, N.Y.)*, *291*(5508), 1560–3.

Fries, P. (2005). A Mechanism for Cognitive Dynamics: Neuronal Communication through Neuronal Coherence. *Trends in Cognitive Sciences*, *9*(10), 474–80.

Gazzaley, A., & Nobre, A. C. (2012). Top-Down Modulation: Bridging Selective Attention and Working Memory. *Trends in Cognitive Sciences*, *16*(2), 129–35.

Harman, G. (1990). The Intrinsic Quality of Experience. *Philosophical Perspectives*, *4*, 31–52.

Johnson, J. S., Hollingworth, A., & Luck, S. J. (2008). The Role of Attention in the Maintenance of Feature Bindings in Visual Short-term Memory. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(1), 41–55.

Kind, A. (2003). What’s So Transparent about Transparency?. *Philosophical Studies*, *115*, 225–244.

Kind, A. (2008). How to Believe in Qualia. In E. Wright (Ed.), *The Case for Qualia*. Cambridge, Mass.: MIT Press.

Kuznetsova, A. Y., & Deth, R. C. (2008). A Model for Modulation of Neuronal Synchronization by D4 Dopamine Receptor-Mediated Phospholipid Methylation. *Journal of Computational Neuroscience*, *24*(3), 314–29.

Lamme, V. A. F. (2003). Why Visual Attention and Awareness are Different. *Trends in Cognitive Sciences*, *7*(1), 12–18.

Loar, B. (2003). Transparent Experience and the Availability of Qualia. In Q. Smith & A. Jokic (Eds.), *Consciousness: New Philosophical Perspectives*. Oxford: Oxford University Press.

Luck, S. J., & Vogel, E. K. (1997). The Capacity of Visual Working Memory for Features and Conjunctions. *Nature*, *390*(6657), 279–81.

Mendelovici, A. (2014). Pure Intentionalism about Moods and Emotions. In U. Kriegel (Ed.), *Current Controversies in Philosophy of Mind*. New York: Routledge.

Papineau, D. (2014). Sensory Experience and Representational Properties. *Proceedings of the Aristotelian Society*, *114*(1).

Prinz, J. J. (2012). *The Conscious Brain*. New York: Oxford University Press.

Treisman, A. (2006). How the Deployment of Attention Determines What We See. *Visual Cognition*, *14*(4-8), 411–443.

Tye, M. (1995). *Ten Problems of Consciousness*. Cambridge, MA: The MIT Press.

Tye, M. (2002). Representationalism and the Transparency of Experience. *Nous*, *36*(1), 137–151.

1. † I am grateful to Hilla Jacobson, Angela Mendelovici and Zohar Bronfman for providing excellent criticisms and suggestions when the core idea of the paper was in its infancy. Thanks to Ori Beck for multiple discussions of the entire manuscript. Thanks also to Amir Horowitz, Takuya Niikawa, John Schwenkler, Miguel Angel Sebastian, Ran Weksler, Wayne Wu, and an anonymous referee for *Philosophical Studies* for helpful feedback. [↑](#footnote-ref-1)
2. I am grateful to Hilla Jacobson for advising me to focus on AS rather than merely on attention to mental paint. [↑](#footnote-ref-2)
3. Block (2003) claims we can attend inwards: he explicitly says that one can “attend to [mental paint] and be aware of it even when one is not attending to it” (p. 173). He also claims we can attend outwards in considering the transparency of experience, “which may be defined as the claim that the effect of concentrating on experience is simply to attend to and be aware of what the experience is of. As a point about attention in one familiar circumstance—for example, looking at a red tomato—this is certainly right.” (p. 171) [↑](#footnote-ref-3)
4. This point was stressed to me by Hilla Jacobson and Zohar Bronfman. [↑](#footnote-ref-4)
5. This claim is slightly inaccurate since Block (2003) distinguishes between mental paint and mental oil. The difference is that the former is representational, as in being a *vehicle* of content (see discussion below), whereas the latter is not. In this paper I focus only on the former. [↑](#footnote-ref-5)
6. Throughout the paper, I will assume physicalism, which I believe helps make the discussion more tractable (for example, it helps clarify the idea of a “vehicle”). The assumption is unproblematic in the present context because the mental paint view is compatible with physicalism, as is evidenced by the fact that Block, Loar, and Papineau are leading defenders of the latter. [↑](#footnote-ref-6)
7. Aside from identifying mental paint with vehicles of experiences, Papineau considers the possibility that mental paint is a *functional* property of experiences. For convenience, throughout the paper, I focus on the former approach, but what I say straightforwardly applies to the latter. [↑](#footnote-ref-7)
8. Prinz does not explicitly say whether he means attending to the content of the representation or to its vehicle. However, he endorses Fries (2001) work, which suggests that, like Fries, he means attending to the content. [↑](#footnote-ref-8)
9. If our experience were to represent mental paint, then the phenomenology of the experience would be exhausted by its representational content, as *intentionalists* hold. [↑](#footnote-ref-9)
10. This assumption appears to be a truism (see Burge’s quote above, my discussion of the neuroscience of attention in Section 5, and Papineau 2014, pp. 26-7), but there is a complication. While Papineau (2014, pp. 26-7) grants that, usually, we are aware of qualities our experience represents, and not of mental paint, he denies this for cases of hallucination or illusion. His point is that we cannot be aware of qualities that are not *instantiated* in our vicinity; we cannot be aware of *merely* represented qualities, in other words. Assuming this point about awareness carries over to attention, it seems that Papineau would deny that, in cases of hallucination and illusion, we can attend to qualities our experience represents. So he would not accept AS for cases of hallucination and illusion, but he would accept it for cases of veridical perception, or so it seems. [↑](#footnote-ref-10)
11. While Burge and Loar clearly say we can attend to mental paint, Kind sometimes says we can be *aware* of mental paint, without explicitly mentioning attention. Papineau never explicitly says we can attend to mental paint, only that we can be *introspectively aware* of it. However, the way Kind and Papineau discuss the matter makes it clear that they are *not* speaking about *awareness* *without attention* (see below). For, they are speaking of a sort of awareness that gives rise to thoughts about mental paint, and arguably we cannot form thoughts about mental paint, on the basis of our awareness of it, without attending to it. [↑](#footnote-ref-11)
12. This claim is slightly inaccurate. During the blank screen stage, one directs attention to the content of a representation *in visual working memory*, not to the content of a conscious visual experience. So it is not accurate to say that, during that stage, one is attending to a quality one’s experience represents. A more accurate formulation is that, during that stage, one directs attention to a feature the experience *had* represented before it disappeared. [↑](#footnote-ref-12)
13. If the reader thinks that (in vision) there is no orientation mental paint (and more generally no spatial mental paint), only color mental paint, then it is possible to modify the experiment so that it will include only colors. Specifically, it will include *conjunctions of colors*, in the form of a square within a square (e.g., a red square within a green square), as in one of Luck and Vogel’s (1997) experiments. The participants will be instructed to attend only to *color* mental paint in all tests, ignoring orientation. [↑](#footnote-ref-13)
14. For a similar formulation of the method see Loar (2003, p. 91). [↑](#footnote-ref-14)
15. Thanks to Takuya Niikawa and Wayne Wu for discussions of Loar’s method. [↑](#footnote-ref-15)
16. Clarification: it *is* possible to form a judgment about more objects than working memory can contain. In that case, however, one keeps in working memory only statistical properties, e.g., average orientation or frequent color (Treisman 2006). Such a statistical memory is of no help in most change detection tasks (depending on their design). [↑](#footnote-ref-16)
17. Thanks to Hilla Jacobson and Angela Mendelovici for raising concerns of broadly this sort. [↑](#footnote-ref-17)
18. I thank Zohar Bronfman for raising a concern along these lines. [↑](#footnote-ref-18)
19. A minor complication: if Prinz is correct in claiming that consciousness requires attention, then VEH1 is red mental paint *only when it is gamma synchronized*. So, strictly speaking, VEH1 is not, in and of itself, red mental paint. Instead, *gamma synchronized* VEH1 is red mental paint. This complication does not matter to the central point I make in what follows. [↑](#footnote-ref-19)