Is Attention Necessary and Sufficient for Phenomenal Consciousness?

Henry Taylor

This is the penultimate version of a paper that appeared in *The Journal of Consciousness Studies* 2013. 20: 173-194

Correspondence: jht30@cam.ac.uk

Abstract.

There has recently been a flurry of interest over how attention and phenomenal consciousness interact. Felipe De Brigard and Jesse Prinz have made the bold claim that attention is necessary and sufficient for phenomenal consciousness. If this turns out to be true, then we will have taken significant steps toward naturalising the mind, which is a particularly exciting prospect.

Against this position, several thinkers have presented empirical data which apparently show that consciousness is possible in the absence of attention, and vice versa. In this paper I argue that these results do not harm De Brigard and Prinz's position, but that this is unsurprising because they use a definition of 'attention' which makes their view empirically self-sealing. I shall also argue that the argument in favour of this definition of attention is unsuccessful. I shall close with some comments on what should be done for the debate to progress. I have three main aims in this paper: firstly to give an overview of the debate, secondly to thoroughly analyse De Brigard and Prinz's position and thirdly (and most importantly) to point out some general and troublesome methodological issues that beset the debate, which have gone largely unacknowledged in the literature. Particularly, I will highlight the cross-purposes which stem from participants using definitions of the key terms in different ways.

1-Conceptual preliminaries.

In this section, I will get clear on how we should understand the claim that 'attention is necessary and sufficient for consciousness'.[1] In sections 2 and 3 I shall outline the empirical data which have been presented against this view, and argue that these data do not harm De Brigard and Prinz's position but that this is unsurprising because they define attention in a tendentious way. I shall discuss only the data which are most prominent in the literature. In section 4 I will argue that their definition of attention makes their view empirically self-sealing. I will consider possible ways of refuting the position, and argue that none of them can work. In sections 2-4 I will not argue that their definition of attention is *false,* I only wish to show how the definition serves to sweep aside all

[1] Prinz has developed the view in more detail than De Brigard, and as a result, many of the claims are made only by Prinz, not both. For this reason, I will attribute such views only to Prinz. It is unclear whether De Birgard would still hold the view that attention is sufficient for consciousness. For example, in his (2012) he claims that attention is insufficient for conscious recollection of memories, and he has informed me that he is also more sceptical of the view that attention is sufficient for *perceptual* representations to become conscious.

possible empirical evidence that could be presented against their position. In section 5 I will consider Prinz's argument in favour of his definition of attention and argue that it is circular. I will also consider an argument Prinz gives which I call the 'threat of eliminativism' and argue that this fails. In section 6 I suggest some steps for resolving these difficulties.

De Brigard and Prinz say this: '[w]e claim that attention is necessary and sufficient for perceptual representations to become conscious' (2010, p.51). We can understand this as a conjunction of the following two claims, which I call NT and ST:

i) *The necessity thesis (NT):*

Attention to some item(s) is necessary for a representation of that item (or those items) to be phenomenally conscious.[2]

ii) *The sufficiency thesis (ST):*

Attention to some item(s) is sufficient for a representation of that item (or those items) to be phenomenally conscious.

ST and NT allow for the possibility that the attention in question may be inward-directed (at our own mental states and events) or outward-directed (at the world). Relatedly, ST and NT leave open the possibility that the item that the subject is attending to may be *self-representing.* That is to say, it may be the case that sometimes (for example, when we pay attention to our own perceptual states) those states represent *themselves* in a phenomenally conscious manner.

We can briefly list the main positions in the debate:[3]

1) Attention is necessary and sufficient for consciousness (De Brigard and Prinz, 2010 and Prinz, 2010, 2011 and 2012).
2) Attention is necessary but not sufficient for consciousness (Cohen et al. 2012).[4]

[2] I take phenomenal consciousness to be a term that applies to all and only those mental events that there is 'something it is like' to undergo (see Nagel, 1974).
[3] This list is obviously not exhaustive.

3) Attention is sufficient but unnecessary for consciousness (Smithies, 2011 and Mole, 2008).[5]

4) Attention is *not* sufficient for consciousness (Kentridge, 2011 Kentridge et al. 1999, 2008a and 2008b and Norman et al. in press).

5) Attention is neither necessary nor sufficient for consciousness (Lamme, 2003, 2010; Koch and Tsuchiya, 2007).[6]

We can now ask why we should care whether or not attention is necessary and sufficient for consciousness. The main reason is that if position (1) were true, then we would have gained an important insight into the nature of phenomenal consciousness, and how it interacts with the rest of the mind. Such an insight would have important implications for philosophy, psychology and neuroscience. Position (1) also opens the door to other important theses about the interaction between attention and consciousness, for example, it may be that one of them causes the other, or that one constitutes the other (in the sense of being a necessary part of it). It could even be that attention and consciousness are identical.[7] Another important reason to care about whether attention is necessary and sufficient for consciousness is that this is the central claim of Prinz's whole 'AIR' theory of consciousness, so it should be examined carefully.

De Brigard and Prinz's arguments in favour of position (1) are based upon inference to the best explanation. They argue that in certain cases, attention to a stimulus can bring it to our consciousness, and that a lack of attention to a stimulus can exclude it from being conscious (see De Brigard and Prinz, 2010, pp.53-54 and Prinz, 2012, pp.81-87). They draw upon cases such as

[4] Tye (2010) defends a thesis similar to NT. De Brigard has told me that his view is now probably closer to (2) than (1).
[5] Though Mole gives a different view in his (2011).
[6] (5) is the most natural reading of Block (2013).
[7] Though such an identity theory can come in various kinds. Prinz himself identifies conscious experiences with mental events that fulfil a certain functional role (being attended to) and then argues that vectorwave firing within the gamma range realises this functional role (2012, ch.9). Prinz, however, thinks that it is a contingent fact that vectorwave firing realises this role, so the identity between consciousness, attention and vectorwave firing does not hold with metaphysical necessity. The general structure of Prinz's view is similar in many ways to that of Lewis (1966).

inattentional blindness, visual pop-out, visual neglect and other phenomena to make this case. From this they conclude that attention is necessary and sufficient for consciousness.

I will not analyse the arguments *in favour* of their central claim. This is firstly because they are not the focus of this paper, and secondly because they have been addressed elsewhere.[8] I should, however, mention that since their argument is mainly based upon an empirical correlation between phenomenal consciousness and attention, and since their claim is a universal one (they claim that attention is *always* necessary and sufficient for phenomenal consciousness) we only need *one* case where either attention occurs without consciousness (disproving ST) or where consciousness occurs in the absence of attention (disproving NT) in order to disprove their claim that attention is necessary and sufficient for consciousness.

2-Supposed evidence against ST.

In this section, I will argue that De Brigard and Prinz's definition of their key terms makes their position impervious to falsification by the empirical results so far put forth. My main focus will be on their definition of 'attention' but I will also raise some important issues concerning their definition of 'working memory' in section 4.2. My aim in this section is not to argue that their view is false but rather to show how their definition of attention serves to insulate their position from falsification. I will assess whether there is good reason to accept their definition of attention in section 5.

2.1-The evidence.

In a series of studies,[9] the subject GY (who suffers from blindsight)[10] was tested in an attempt to determine whether he could attend to stimuli in his blind field. The experimenters set up a test where they would present target stimuli in the blind area of GY's visual field. Since the

[8] E.g. Mole (2011, ch.7).
[9] See Kentridge et al. (1999, 2008a and 2008b). See also Norman et al. (in press) and Kentridge, (2011).
[10] Blindsight is a condition where subjects deny awareness of items placed in certain areas of their visual field, but are susceptible to priming and other subliminal effects which show that information about the items is processed in their visual system.

stimulus was in GY's blind area, he denied seeing it. A tone was sounded on occasion, sometimes coinciding with the presentation of the target stimulus in GY's blind area, sometimes not. GY was encouraged to respond as quickly as possible after the tone was sounded *if* he felt that the target had been presented in his blind field. It was found that GY was quicker to respond to the target stimulus if the location of the stimulus was prior indicated by the use of a cue. The cue was presented in the healthy area of GY's visual field.[11] The experimenters concluded that GY was paying attention to the target stimulus, even though he had no phenomenal representation of it: 'attention could selectively modulate the processing of a target without that target's entering awareness' (2011, p.240).

Following this, studies were performed in an attempt to demonstrate attention to certain items in the absence of phenomenal consciousness of the attended items in non-neurologically impaired subjects. It was found that these subjects were more likely to be primed by an unconscious stimulus when arrows (which were presented in a location visible to the subjects) pointed toward that stimulus (see Kentridge et al. 2008b and Kentridge, 2011, p.240). [12]

2.2-Why the results do not damage the position.

We can see how De Brigard and Prinz's position is not damaged by these results when we look at how they define attention. Here is Prinz's definition: 'attention can be identified with the processes that allow information to be encoded in working memory. When a stimulus is attended, it becomes available to working memory, and if it is unattended, it is unavailable' (Prinz, 2011, p.184. cf. Prinz, 2012, p.93 and De Brigard and Prinz, 2010, p.52).

[11] In some manipulations, cues were used which fell in GY's blind field (in these manipulations, he was aware of neither the cue nor the target). His reaction speed also increased in these manipulations (see Kentridge et al., 1999, pp.1805ff.). These peripheral cues were pairs of bars presented around the location where the target would appear.

[12] The stimuli underwent meta-contrast masking to ensure that the subjects were not phenomenally conscious of them. Masking is a process where the outer edges of an initial stimulus (the 'masked' stimulus) coincide with the inner edges of a subsequently presented 'mask' stimulus. The result is for signals from the masked stimulus to coincide with signals from the mask in the early stages of processing in the visual cortex, resulting in the masked stimulus not being consciously perceived.

So, attention is to be identified with the faculty that makes information available to working memory, all and only information that is available to working memory is attended to. How do they understand 'working memory'?

Working memory, as Prinz understands it, is 'a short-term storage capacity, but one that allows for "executive control"… Once something is encoded in working memory, it becomes available to language systems for reporting, and with systems that allow effortful serial processing' (2011, p.184). Prinz also expands the idea thus: '[t]he attended stimulus becomes available for processes that are controlled and deliberative. For example, we can *report* the stimulus that we consciously perceive, we can reason about it, we can keep it in our minds for a while, and we can wilfully choose to examine it further' (2012, p.92).

The important thing to note about these definitions of working memory is the importance of *reportability.* That is, Prinz notes that one of the functions of working memory is to make certain information reportable. Information which is not available to working memory is not reportable; information which is available to working memory is reportable. Notice also that Prinz mentions that information available to working memory can (by definition) be used for *controlled and deliberative* action, so if a subject shows priming affects after being exposed to a stimulus, i.e. if she can report the stimulus when given a forced choice, but denies seeing it, and claims only to be guessing what was there, then the stimulus does not count as having entered working memory.

With this definition of attention, we can now see how the studies in question do not harm ST. The important detail lies in the way that the experimenters can establish that the subject did not have a phenomenal representation of the stimulus they were (supposedly) attending to. In order to establish this, the experimenters had to *ask* the subjects whether they saw the stimulus. The subjects in question, of course, denied that they saw the stimulus. This is taken as evidence that they were not phenomenally conscious of it (Kentridge, 2011, p.230). Once this has been established, the

experimenters attempt to establish that the subjects *paid attention* to the stimuli that they denied seeing. If they can establish both, then we have evidence against ST.

The problem is that *if* the subjects do indeed deny seeing the stimulus (which is how the experimenters established that the stimulus was unconscious), then it will follow that the stimulus was not available to working memory, and thus (by De Brigard and Prinz's *definition* of attention) that it was not attended. So immediately the data will not count as an instance of attention to something which was not phenomenally conscious. So the data in question will not count as a counterexample to ST.

Indeed, Prinz seems to make just such an argument (though he does not state it in such stark terms) when he says the following: '[i]f my earlier analysis of attention is right, attention entails availability to working memory. Availability is clearly absent in blindsight, so GY cannot be instantiating *all* of the processes necessary for attention' (2012, p.115 cf. 2011, p.194). Prinz relies upon his earlier definition of attention in order to here reject the Kentridge studies as showing an instance of attention without phenomenal consciousness.

What we see, then, is that De Brigard and Prinz's position is completely immune to falsification by the results that Kentridge et al. put forth. Ordinarily, of course, this would be good news for their position. However, there is clearly something suspect here. The problem is that the work is being done by their *definition* of attention. De Brigard and Prinz's position derives a substantial amount of its force from the way that they define their terms.

2.3-Orienting.

I have argued that De Brigard's and Prinz's definition of attention allows them to dismiss the evidence presented against ST. However, there is another route that De Brigard and Prinz could take, which is to attempt to explain the data without committing themselves to the claim that GY is attending to something of which he is not conscious *without* relying on their tendentious definition

of attention. Indeed, such a response has been given (see Prinz, 2011, pp.193-196 and 2012, pp.113-118). Prinz suggests that GY's performance could be due to GY *orienting* to the target, rather than actually *attending* to it. [13]

There are two things that I will say with respect to this possible response. The first is that there is now reasonably good empirical evidence that the orienting response is untenable. Unfortunately, I cannot give full details here, but the main issue is that orienting responses do nothing to favour the processing of a target which occurs *within* the same object as the cue, but this preferential processing is just what we find in some of the experiments in question. There is a related point to be made here that it is questionable whether the items in question could trigger an orienting response, due to their extremely low salience.[14]

Secondly, whether the orienting response is viable or not, it will not affect the main point that I wish to make, which is that it *does not matter* exactly whether such responses are plausible, because De Brigard and Prinz's position can already dismiss the empirical evidence presented against their view, simply in virtue of their definition of attention. The main problem I wish to show is that *no matter* how the empirical evidence turns out, De Brigard and Prinz's *definition* of attention will always be sufficiently tendentious to brush aside the GY results as not really instances of attention without consciousness. The definition of attention supplied will always serve to protect the position from the results in question.

In this section I have argued that De Brigard and Prinz's definition of attention makes their position impervious to falsification by the results in question. In section 4 I will argue that their definition of attention makes their position immune from falsification from *any* possible empirical results, but first we must consider the evidence put forth against NT.

[13] 'Orienting' is the term Prinz gives to the collection of processes which control what information enters the visual system (2012, pp.113-4). Note that the results *cannot* be explained in terms of overt movements of the eye toward the target, because eye movement was fixed in the experiment (this was verified with an eye-tracker). Rather, the claim is that the neurons that respond to the area of the visual field where the target was presented become more *sensitive*, and that this increase in sensitivity can explain subjects' performance.
[14] For a more detailed explanation of these data see Norman et al. (in press) and McCarley et al. (2002).

3-Supposed evidence against NT.

3.1-The evidence.

I have argued that the empirical data against ST are unable to harm De Brigard and Prinz's position, due to their definition of attention. I shall now show how a similar situation arises when we are considering the data that has been put forward against NT.

The most prominent argument against the claim that attention is necessary for consciousness is based on empirical results by Li et al. (2002). In these experiments, subjects are asked to perform a task which requires a lot of attention (they had to work out whether a collection of letters contained an 'L' or not) and at the same time as they were performing this task, an image of something was flashed up in an area outside the focus of their attention. Subjects were asked to release a button when they detected a target stimulus in the peripheral location (i.e. outside of the focus of attention). It was found that subjects could report the gist of the image flashed up even though they were concentrating their attention on the letter identifying task.[15] The authors of the article themselves take this to be an instance of phenomenally conscious experience in the near absence of attention, though others[16] have held it to be an instance of phenomenal consciousness in the *complete* absence of attention. It is these results which have most prominently been taken as evidence against NT.

The core feature about these experiments that we must focus upon, is that in addition to demonstrating that attention to the peripheral stimulus was absent (or nearly absent), the experimenters must also demonstrate that the subjects had a phenomenal experience of the peripheral stimulus. In order to do this, of course, they had to *ask* the subject whether they could

[15] This was a 'free report' task, making it very unlikely that the results can be attributed to unconscious priming of the subjects by the peripheral targets (cf. Azzopardi and Cowey, 1997).
[16] Ned Block, in conversation.

see the stimulus. The reportability of the stimulus is taken as evidence that the subjects had a phenomenal representation of the stimulus.

3.2-Why the evidence does not damage the position.

Just as was the case with ST, these results do not harm the position of Prinz and De Brigard. The crucial point is that the subjects were able to *report* the presence of the image that was flashed up outside the focus of attention. So it must follow that the image was *attended to* (because only representations available to working memory are reportable in this sense, and all and only representations available to working memory are attended, according to De Brigard and Prinz). So, the Li et al. results do not count as an instance of consciousness without attention, in virtue of their definition of attention.

As was the case with ST, it may prove possible to accommodate these data by turning to other interpretations that do not rely explicitly on De Brigard and Prinz's definition of attention. One option (e.g. Cohen et al. 2012 and Cohen and Dennett, 2011) is to claim that the peripheral stimuli may have been subject to a kind of 'distributed' rather than focal attention, and thus the Li et al. results do not count as an instance of consciousness in the absence of attention. However, as before, even if this were plausible, this would not affect my main point, which is that it *does not matter* whether such responses are viable, because De Brigard and Prinz's position is invulnerable to falsification from these results, simply in virtue of their definition of attention.

4-Could De Brigard and Prinz's claims be disproven?

Here I shall argue that De Brigard and Prinz's claims are impervious not only to falsification by the *available* evidence but that it is hard to see how *any* evidence could possibly falsify their position. I am not arguing that their definition is *false* (though I shall later argue against the argument given in favour of their definition) I am arguing that their definition makes their position

empirically self-sealing. In this section I shall also examine a related issue, which is a set of difficulties arising from Prinz's definition of 'working memory'.

4.1-Is the position empirically self-sealing?

The problems elucidated in sections 2 and 3 are symptomatic of a deeper problem, which is that De Brigard and Prinz's position appears to be entirely impervious to empirical falsification. To see this, consider the following argument:

1) In experimental settings involving human subjects[17] reportability is always used in order to establish phenomenal consciousness. That is to say, if subjects can report the presence of a stimulus then they are taken to be conscious of it, and if they cannot report it, they are taken not to be conscious of it.
2) In order to disprove NT or ST, we would require a case where attention and phenomenal consciousness dissociate.
3) De Brigard and Prinz have linked their account of attention with reportability (by defining it in terms of working memory).
4) (Therefore) anything which we can establish is conscious will also count as attended to, and anything that we can establish as unconscious will count as unattended to, as defined by De Brigard and Prinz.
5) (Therefore) Their position cannot be empirically disproved.

In order to support this argument, consider what a counterexample to their position might look like. What would be required to disprove ST is a subject who said that they could see something, freely report it, and use information about it to guide their action in a wilful and deliberate manner, and interact with it just as a normal human can, but who still denies having any phenomenal consciousness of that thing. This would be an example of something which was available to working

[17] I include this qualifier in order to distinguish between consciousness studies on humans from studies on monkeys, which do not utilise verbal reports (see Logothetis and Schall, 1989).

memory (and thus attended to) but which was not phenomenally conscious, and would thus serve as a counterexample to ST.

Unfortunately, this clearly sets the bar too high. What we would effectively be asking for is something approaching a philosophical zombie. Not even an epiphenomenalist would be likely claim that such a zombie was physically possible.[18] Indeed, not only would it clearly be asking too much to request such extravagant empirical evidence against ST, but even if we did have such a subject, then we would be far more likely *not* to believe her claim that she was not phenomenally conscious of the item in question. Rather, we would probably conclude that she was lying, or deluded.[19]

Similar things go for any possible counterexamples to NT. In order to demonstrate a counterexample to NT, what we would require for NT to be falsified is a case of a conscious experience which the subject in question was totally unaware of, and actually *denied having.* Again, such a case is extremely fanciful, and too much to hope for. It is hard to see how we could ever establish that there existed such a phenomenal experience, given that subjects would actually deny having it.[20]

4.2-Unconscious working memory?

It may be argued that there already exists an empirical example of information that is available to working memory, but which is unconscious.[21] If this were true, then it would serve as an example of something which is attended to (by the definition of attention at issue) but which was unconscious, thus disproving ST. The relevant example comes from Soto et al. (2011) who established that subjects can perform above chance at a task comparing the orientation of a subliminally presented Gabor patch (of which the subjects were not conscious) to a supraliminally

[18] I say 'physically possible' to set the modal strength of the claim aside from 'metaphysical possibility', which is what is of import in the zombie debates.
[19] Cf. Dennett (1995).
[20] Though see Lamme (2010).
[21] Thanks to an anonymous referee for emphasising this.

presented Gabor patch (of which the subjects were conscious).[22] The experimenters conclude that the subliminally presented patch was encoded in working memory (and thus must have been *available* to working memory) but was unconscious. If this were true, it would clearly be intolerable to De Brigard and Prinz's position. Are the Soto et al. results a counterexample to their claim?

The answer is no, and the reason once again lies with definitions. In order to see this, we will need to step away from the main issue of this paper, which is the use of the term 'attention' and consider the use of some of the other terms in the debate, specifically the use of the term 'working memory'. As I will now argue, there are similar issues here to those that I have explained above.

Prinz's response to the Soto et al. data is to claim that the subliminal Gabor patch was not really encoded in working memory (2012, p.96). When we examine the definitions of working memory at issue, we can see why this is. Recall that when Prinz defines working memory, he mentions reportability and explicitly says that if something is unreportable, it must be unavailable to working memory. In the Soto et al. task, the experimenters had to *ask* whether the subjects saw the subliminally presented Gabor patch, and rate their awareness of it on a scale (2011, R912). If subjects reported that they had 'no awareness' of the subliminally presented Gabor patch, the experimenters concluded that it was unconscious. Now, the problem will be obvious. If the subjects deny seeing the Gabor patch, then it will follow from Prinz's *definition* of working memory (which includes reportability) that it was not available to working memory. So the subliminally presented Gabor patches will not count as an instance of unconscious representations that are available to working memory, and thus not an example of something unconscious but attended, and thus not a counterexample to ST. Notice that here, much of the work is being done by Prinz's definition of working memory, which in turn affects his definition of attention.

4.3-The neural correlates of working memory?

[22] A Gabor patch is a rippled texture, tilted to a specific orientation.

Felipe De Brigard has suggested[23] that there is a hypothetical experiment which *could* disprove De Brigard and Prinz's claim. In this hypothetical experiment, one group of subjects would attend to a stimulus, and (presumably) report awareness of it. Then a second group would perform the same task whilst undergoing some kind of brain manipulation which interrupts the neural circuits which Prinz claims underpin attention. If the second group still report seeing the stimulus, this will be evidence against the position.

However, this will not disprove NT, because the same issue as we encountered above will reemerge, which is that *if* the subjects in the second manipulation report seeing the stimuli, then those stimuli will automatically count as available to working memory *regardless of the neural details* and thus they will count as attended to, and so the experiment will not count as a counterexample to the position. Prinz's definition of attention is functional, so if subjects fulfil that functional role (by being able to report the stimuli) then they will count as attending to the stimuli, *no matter what* is happening in their brains.

It may be replied that such an experiment could disprove *some* aspects of Prinz's theory, however. Specifically, Prinz claims that what he calls or 'gamma-locked oscillations' or 'gamma synchrony' are the neural entities that fulfil the role associated with attention. It may be said that if it transpires that the subjects in the second manipulation still could report the presence of the stimuli, despite the brain manipulation interrupting the relevant brain properties, then this disproves Prinz's claim about the neural correlates of attention being gamma synchrony.[24]

However, here we must be careful. Such evidence would disprove a part of Prinz's theory, but it would not disprove the claims that we are really interested in, which are NT and ST. To see this, we can deconstruct Prinz's theory into several different claims:

i) Attention is necessary and sufficient for consciousness.

[23] Personal communication.
[24] Thanks to James Stazicker for suggesting this to me.

ii) Attention is defined as the process that underpins availability to working memory.

iii) Working memory should be defined a certain way.

iv) Gamma synchrony realises the role associated with attention.

Now, the experiment outlined above may be able to disprove (iv), but it could not affect the claims that this paper is primarily concerned with, which are (i-iii), because they can all be true independently of (iv). So if it did indeed transpire that availability to working memory can dissociate from gamma synchrony, then all that would follow would be that availability to working memory is *not* in fact realised by gamma synchrony, but this would only show that we need to find some different brain properties that fulfil the role of availability to working memory. The claim that attention (defined functionally) should be identified with availability to working memory and the related claim that NT and ST are true would not themselves be damaged by this evidence, and it is after all these claims that we are interested in. Relatedly, all of the functional definitions of 'attention' and 'working memory' which make De Brigard and Prinz's version of NT and ST self-sealing would still stand.

I should emphasise that I am not arguing that *all* aspects of Prinz's rich and varied theory of consciousness are empirically self-sealing, I only claim that his version of NT and ST are empirically self-sealing, and that this is primarily due to how he defines 'attention', other aspects of his theory (such as which properties of the brain he thinks realise the role of availability to working memory) could prove false, but this would not endanger the main theses that we have been discussing.

In summary, the basic problem is that Prinz does not allow attention, reportability and availability to working memory to dissociate, and also takes reportability as evidence of phenomenal consciousness. For this reason, it is almost analytic to claim that there is evidence of phenomenal consciousness when there is evidence of attention, and it is unsurprising that all proposed counterevidence to the claim does not hit the mark.

5-Is there good reason to accept De Brigard and Prinz's definition of attention?

Perhaps we would be willing to accept these problems if there were a good *argument* which showed that De Brigard and Prinz's definition of attention was independently plausible. De Brigard and Prinz do put forth such an argument,[25] but Prinz (2012) has worked it out in the most detail, so I shall concentrate upon his formulation of it. In section 5.1 I shall argue that Prinz's argument is circular. In section 5.2 I shall consider another argument that Prinz gives, and argue against that.

5.1-Prinz's argument.

Prinz proposes that we list 'paradigm' instances of attention and then attempt to discover whether there is a common brain mechanism that underlies them all. If we find such a mechanism, we can identify it with attention. Prinz then goes on to list some instances of cases where attention seems to make information available to working memory (2012, pp.90-95). These include studies which link attention with short term memory retention (Rock and Gutman (1981)) as well as studies which show that when working memory is full, it becomes harder to attend (Fougnie and Marois, 2007).

Prinz then makes the following claim: '[s]uch interactions between attention and working memory suggest an intimate relationship. The simplest explanation for this relationship is an identity claim: attention can be identified with the processes that allow information to be encoded in working memory' (2012, p.93). Prinz then goes on to claim that '[t]he idea of availability underlies all of the phenomena that we call attention' (2012, p.95) and that 'the folk-psychological insight implicit in the range of phenomena that we call attention can map onto the empirical construct of availability to working memory' (2012, p.95).[26]

[25] See De Brigard and Prinz (2010, pp.51-53).

[26] Prinz seems to rely on the folk psychological concept of attention, and then attempts to find the physical entity in the brain which fulfils the roles associated with that concept. His method therefore has a strong resemblance to the 'Ramsification' method made famous by Lewis (1966 and 1970).

The core premise of Prinz's argument in favour of his definition is that attention and availability to working memory always coincide in folk psychological discourse.[27] However, Prinz does not consider all of the borderline cases where attention appears to occur in the absence of availability to working memory. For example, in Ronald Rensink's taxonomy of attention (2013), one form of (visual) attention that is listed is what Rensink calls 'sampling' which is 'the pickup of information by the eye'. Another is 'filtering' which determines which information that the eye receives then travels further along the visual system for processing. Both of these processes occur before information becomes available to working memory (Rensink himself claims that filtering is operative in cases of subliminal perception (2013, §1.4ii)).

Now, here we have a case where a proposed instance of attention occurs in the absence of availability to working memory, how can Prinz respond to such a proposed example? Presumably, Prinz will have to insist that these are not 'really' instances of attention, but what reason (other than merely saving his theory) could we have to make such a postulation?

There is the threat of circularity in the offing. In order to argue that we should define attention in terms of availability to working memory, Prinz needs to argue that attention and availability to working memory always coincide, but in order to do this, Prinz must claim that processes such as 'filtering' and 'sampling' are not really instances of attention (because they can occur in the absence of availability to working memory). But it is hard to see what good reason we might have for thinking that they are not 'really' instances of attention, unless we were already convinced that attention cannot occur in the absence of availability to working memory. But that is precisely the conclusion that Prinz's argument was supposed to show, so Prinz's argument comes close to being circular.

[27] De Birgard (2010) argues that the folk psychological concept of attention is in fact imperfectly delineated. This may be another possible avenue of attack against Prinz.

One possible avenue that Prinz may pursue in order to argue that 'sampling' and 'filtering' are not really instances of attention is to claim that we do not think of such things as 'attention' in our folk psychological discourse. However, this seems unclear. Most folk psychological speakers are unaware of the subtle distinctions between processes such as 'sampling' and 'filtering', and so it is very unlikely that normal naïve subjects will have clear cut intuitions about them one way or the other. Prinz would be on extremely shaky ground attempting to argue that normal naïve folk had a unified view on such matters.

I should say that I am not claiming that such borderline cases certainly *do* count as instances of attention, all I am saying is that Prinz's argument will not go through unless he assumes that they do not, but there seems little reason to assume that they do not unless we are already convinced that Prinz's definition of attention is correct, which is what Prinz's argument was intended to show in the first place.

But what of the empirical results that Prinz cites? Do they not show that there is an intimate relationship between working memory and attention? In response to this, I think we can accept that attention and working memory often interact closely (no one would deny this), but we need not commit to the conclusion that attention *must* be *identified* with availability to working memory. For these reason, I conclude that Prinz's argument in favour of his definition of attention is inconclusive.

A further issue in the vicinity is that if we do accept that attention and availability to working memory should be identified, then all of the literature surrounding the question of how attention and working memory interact (e.g. Fougnie, 2008)[28] would become trivial. If we are to accept the claim that there are substantive questions about how working memory and attention interact, and that there are substantial empirical discoveries that can be made and have been made on this question, then Prinz's definition of attention will not do, because Prinz's definition of attention delivers a simple definitional identity between the two processes.

[28] Watzl (2011) makes a related point.

5.2-The threat of eliminativism.

There is a final argument that Prinz gives, which may be thought to undermine some of my claims. Prinz says this:

> '[t]here may be a common denominator [which applies to all and only instances of attention] that can be empirically discovered. If such a common mechanism were found, we might say that "attention" refers to that mechanism. If these phenomena share nothing in common, then we might say that "attention" should be dropped as a term from scientific psychology. We might become eliminativists' (2012, p.91).

Prinz is here expressing a view that is held by other thinkers on attention. Smithies (2011, p.251) similarly claims that if we do not find a 'unique locus of attentional selection' then this may 'yield a form of eliminativism'. Equally, Allport (1993, p.203) claims that, because 'there is no one uniform computational function, or mental operation in general' that we can identify with attention, then '*there can be no such thing as attention'.*

Prinz may use this point to claim that if we weigh up the available options, it is better that we accept his own (albeit tendentious) definition of attention, rather than become eliminativists about it. Perhaps Prinz's view is the lesser of two evils.

I am unconvinced by this appeal to the threat of eliminativism. In response to it, I will say three things. Firstly, I should point out that I am not claiming that there is *no* common mechanism in the brain that we should identify attention with, just that Prinz's argument is insufficient to demonstrate that attention should be identified with availability to working memory.

Secondly, it is not even clear that Prinz *should* be concerned if attention turns out to be underwritten by many different systems in the brain.[29] To see this, note that Prinz defines attention functionally, so not finding a common 'mechanism' that underwrites all forms of attention should

[29] Thanks to Felipe De Brigard for pressing me on this.

not worry us. So long as all of these mechanisms fulfil the functional roles that Prinz associates with attention, then they will all count as 'attention', using the functional analysis in question.

Thirdly, and most importantly, I think the claim that if we do not find one common mechanism in the brain that we can identify attention with then we will be threatened with eliminativism is very implausible. In order to see how we can avoid eliminativism, consider an analogous case, which is that of memory. [30]

In psychological study of memory the concept of 'memory' will likely be divided into different subsystems, such as episodic memory, declarative memory, non-declarative memory, long term memory, working memory, iconic memory and so on. [31] We know that in some cases, different kinds of memory operate relatively independently of each other, using different mechanisms and operating in different ways. For example, one form of iconic memory operates in the retina, due to the fact that retinal cells continue to fire briefly after the eyelid has closed (Long and Sakitt, 1980 and Block, 2011, p.571). Conversely, certain kinds of long term memory operate in the hippocampus, using long term potentiation of synapses in order to store information for recall.

'Memory' is thus a psychological capacity, underwritten by many heterogenous subsystems, many of which do not share 'common mechanisms', but it would be very strange to claim that this implies that memory does not exist, and it is not as though 'memory' has been eliminated from our psychological discourse. If it does transpire that there is no one 'common mechanism' that can be identified with attention then attention will stand in no worse position ontologically than memory does now. It simply does not follow from the fact that there is no common mechanism that we can identify attention with that attention does not exist. It may well be that our folk psychological concept of attention does not match up perfectly with one mechanism in the brain, but this should

[30] Chun et al. (2011) and De Brigard (2012) also liken attention to memory.
[31] See e.g. Baddeley et al (2009). See also Sligte et al. (2008 and 2009) for a recent bifurcation in the concept of short term visual memory.

not surprise us, that is what folk psychological predicates are like, and we certainly do not need to leap to eliminativism if that is the case.

Another point that De Brigard and Prinz may make is that what I have been discussing is merely a verbal issue, over how to define 'attention'. It may be claimed that the thinkers in this field are simply working with different ideas of what 'attention' is, but that this is merely a linguistic issue, not a substantive one.

In response to this, I claim that if the debate in question is to have any real substance, if there really is to be a definitive answer to the question of whether attention is necessary and sufficient for consciousness, then obviously we are not free to define 'attention' in any way that we choose. If we simply claim that the theorists in question are just working with different concepts of attention, and that this is simply a linguistic issue, then we are dangerously close to saying that 'really' there is no answer to the question of whether attention is necessary and sufficient for consciousness, because different theorists will deliver different answers depending upon their definition of 'attention'. To make this claim is to concede that the question of whether attention is necessary and sufficient for consciousness is itself merely verbal, which is essentially to give up on the whole debate.

6-What to do.

I have been elucidating how apparently innocuous definitions of terms such as 'attention' (and also 'working memory') can have substantial (though usually unacknowledged) force in the debates over ST and NT. In this section I will briefly outline two possible routes that we may take to make progress on these issues.

The first would be to shift the focus of the debate away from the question of whether ST or NT is true, and focus instead on the nature of attention itself.[32] The idea would be that we should

[32] See Watzl (2011) for a survey of this issue.

reflect upon our concept of attention and attempt to come up with a unified account of what attention is. Only then can we assess whether attention is necessary and sufficient for consciousness.

This option may be viable, but it relies upon a heavy assumption, which is that we will be able to formulate such a unifying account. There are several reasons that we may be sceptical of this. Firstly, we may not put much faith in conceptual analysis at all (e.g. Quine, 1951). Secondly we may think that the word 'attention' most likely covers a heterogenous range of different phenomena, which makes the project of obtaining a unifying account of all of them appear quixotic (cf. Duncan, 2006).[33]

Another approach (suggested by De Brigard himself (2010, p.200)) is more amenable to the approach of empirical psychology, which is that we develop operationalist definitions of the phenomena in question, designed to make ST and NT empirically testable. Such definitions may well be tailored to specific experimental paradigms. Obviously, it will be of central importance that one uniform definition of the phenomena be used by different interlocutors in the debate. The advantages of this approach in making the questions empirically tractable are obvious, though it seems likely that such a definition will depart from a normal folk psychological understanding of attention, in which case we will have to ask serious questions about whether philosophers and empirical psychologists are really talking about the same thing. In any case, one thing that seems likely is that De Brigard and Prinz's definition will not serve these empirically-focussed purposes, as their definition appears self-sealing.

7-Summary.

I have argued that De Brigard and Prinz's claim that attention is necessary and sufficient for consciousness is virtually impervious to any empirical falsification, but that this is due to their

[33] This is my own view.

definition of attention. I have considered the reasons given for accepting their definition, and rejected them.

My main concern in this paper was to highlight the difficulties which often go unacknowledged in these debates, which are caused by the definitions of the terms involved. I hope to have shown at the very least that we need to be more careful when approaching these issues, because of the problems embodied in our understanding of terms such as 'attention' and also terms such as 'working memory'. One thing that is clear is that progress in this debate seems unlikely if we do not pay more attention to the definitions of key terms involved.[34]

References.

Allport, A. (1993) Attention and Control. Have we been asking the wrong questions? A critical review of twenty-five years. In Meyer, D. E. and Kornblum, S. (eds.) *Attention and Performance XIV.*(Cambridge, MA: MIT Press.

Azzapardi, P. and Cowey, A. 1997. Is blindsight like normal, near-threshold vision? *PNAS* **94**, pp.14190-14194.

Baddeley, A., Eysenck, M. and Anderson, M. (2009) *Memory.* New York: Psychology Press. Reprinted 2010.

Block, N. (2011) Perceptual consciousness overflows cognitive access, *Trends In Cognitive Sciences.* **15** (12), pp.567-575.

Block, N. (2013) The Grain of Vision and the Grain of Attention. *Thought: A Journal of Philosophy.* doi: 10.1002/tht3.28.

Chun, M., Golomb, J. and Turk-Browne, N.B. 2011. A taxonomy of internal and external attention. *Annual Review of Psychology.* 62, 73-101.

Cohen, M. and Dennett, D. (2011) Consciousness cannot be separated from function, *Trends In Cognitive Sciences,* 15 (8), pp.358-364.

Cohen, M.A., Cavanagh, P., Chun, M. and Nakayama, K. (2012) The attentional requirements of consciousness, *Trends In Cognitive Sciences.* **16** (8), pp. 411-417.

[34] Thanks to Felipe De Brigard, E. J. Lowe and an anonymous referee for comments on previous drafts of the paper. Thanks also to Ned Block and James Stazicker for interesting discussion. Special thanks to Bob Kentridge for many patient discussions while the paper was being written.

De Brigard, F. (2010) Consciousness, attention and commonsense. *Journal of Consciousness Studies.* **17** (9-10), pp.189-201.

De Brigard, F. (2012) The role of attention in conscious recollection. *Frontiers in Psychology.* **3** pp.1-10.

De Brigard, F. and Prinz, J. (2010) Attention and consciousness. *Wiley interdiscerplinary reviews: Cognitive science.* **1** (1), pp.51-59.

Dennett, D. C. (1995) The Path not Taken. Reprinted in (1997) Block, N., Gulzeldere, G. and Flanagan, O. (eds.) *The Nature of Consciousness: Philosophical Debates.* USA: MIT Press.

Duncan, John. (2006) Brain mechanisms of attention. *The Quarterly Journal of Experimental Psychology.* **59**, pp.2-27.

Fougnie, D. (2008) The relationship between attention and working memory. In. Johansen, N. B. (ed.) *New Research on Short Term Memory.* (New York: Nova Science Publishers).

Fougnie, D. and Marois, R. (2007) Executive load in working memory induces inattentional blindness. *Psychonomic Bulletin and Review.* **14** (1) pp.142-147.

Kentridge, R. (2011) Attention without awareness: a brief review. In Mole, C., Wu, W. and Smithies, D. (eds.) *Attention: Philosophical and Psychological Essays.* New York: Oxford University Press.

Kentridge, R., Heywood, C.A. and Weiskrantz, L. (1999) Attention without awareness in blindsight. *Proceedings of the Royal Society (London) Series B.* **266** pp.1805-1811.

Kentridge, R., de-Wit, L.H. and Heywood, C.A. (2008a) What is attended in spatial attention? *Journal of Consciousness Studies.* **15** (4) pp.105-111.

Kentridge, R., Nijober, T. C. W. and Heywood, C. A. (2008b) Attended but unseen: Visual attention is not sufficient for visual awareness. *Neuropsychologia.* **46** (3) pp.831-69.

Koch, C. and Tsuchiya, N. (2007) Consciousness and Attention: Two distinct brain processes. *Trends in Cognitive Sciences.* **11** (1), pp.16-22.

Lamme, V. (2003) Why visual attention and awareness are different, *Trends in Cognitive Sciences.* **7** (1), pp.12-18.

Lamme, V. (2010) How neuroscience will change our view on consciousness, *Cognitive Neuroscience.* **1** (3), pp.204-240.

Lewis, D. (1966) An argument for the identity theory, *The Journal of Philosophy,* **63** (1), pp.17-25.

Lewis, D. (1970) How to define theoretical terms, *The Journal of Philosophy,* **67** (13), pp.427-446.

Li, F. F., VanRullen, R., Koch, C. and Perona, P. (2002) Rapid Natural Scene Categorization in the Near Absence of Attention, *Proceedings of the National Academy of Sciences of the United States of America.* **99** (14), pp.9596-9601.

Logothetis, N. and Schall, J. (1989) Neural correlates of subjective visual perception, *Science,* **245** (4919), pp.761-763.

Long, G. M. and Sakitt, B. (1980) The retinal basis of iconic memory: Eriksen and Collins Revisited, *American Journal of Psychology.* **93** (2), pp.195-206.

McCarley, J., Kramer, A. and Peterson, M. (2002). Overt and covert object-based attention. *Psychonomic Bulletin and Review.* **9** (4), pp.751-758.

Mole, C. (2008) Attention and consciousness. *Journal of Consciousness Studies.* **15** (4), pp.86-104.

Mole, C. (2011) *Attention is Cognitive Unison: an essay in philosophical psychology.* New York: Oxford University Press.

Nagel, T. (1974) What is it like to be a Bat? Reprinted in (2004) Heil, J. (ed.) *Philosophy of Mind: A Guide and Anthology.* New York: Oxford University Press.

Norman, L.J., Heywood, C.A. & Kentridge, R.W. (in press) Object-based attention without awareness. to appear in *Psychological Science.*

Prinz, J. (2010) When is perception conscious? In Nanay, B. (ed.) *Perceiving the World.* New York: Oxford University Press.

Prinz, J. (2011) Is attention necessary and sufficient for consciousness? In Mole, C., Wu, W. and Smithies, D. (eds.) *Attention: Philosophical and Psychological Essays.* New York: Oxford University Press.

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

Quine, W.V.O. 1951. Two dogmas of empiricism. In *From a Logical Point of View: 9 Logico-Philosophical Essays.* USA: Harvard. Reprinted 1981.

Rensink, R. 2013. Perception and attention. In Reisberg, D. (ed.) *The Oxford Handbook of Cognitive Psychology.*

Rock, I and Gutman, D. (1981) The effect of inattention on form perception. *Journal of Experimental Psychology: Human Perception and Performance.* **7** (2), pp.275-285.

Sligte I.G., Scholte H.S., Lamme V.A.F. (2008) Are There Multiple Visual Short-Term Memory Stores? *PLoS ONE* **3** (2): e1699. doi:10.1371/journal.pone.0001699 [09/01/2013].

Sligte, I.G., Scholte, S. and Lamme, V.A.F. (2009) V4 Activity Predicts the strength of Visual Short term memory representations. *The Journal of Neuroscience.* **29** (23), pp.7432-7438.

Smithies, D. (2011) Attention is rational access-consciousness. In Mole, C., Wu, W. and Smithies, D. (eds.) *Attention: Philosophical and Psychological Essays.* New York: Oxford University Press.

Soto, D., Mäntylän T. and Silvanto, J. (2011) Working memory without consciousness. *Current Biology.* **21** (22), pp.R912-R193.

Tye, M. (2010) Attention, seeing and change blindness. *Philosophical Issues.* **20** (1), pp.410-437.

Watzl, S. (2011) The nature of attention, *Philosophy Compass.* **6** (10), pp.842-853.