SCHOOL *of* PHILOSOPHY, PSYCHOLOGY
*and* LANGUAGE SCIENCES

The University of Edinburgh
Dugald Stewart Building
George Square
Edinburgh EH8 9AD

Switchboard +44 (0) 131 650 1000
Fax +44 (0) 131 650 3660
http://www.ppls.ed.ac.uk

**Re: Report on 'Naturalizing the Unity of Consciousness: can neuroscience explain a fundamental feature of subjectivity?**

This is an impressively well-informed dissertation that does a superb job of marshalling, explaining and discussing a range of technical material from different disciplines in a way that makes progress on a set of fundamental but neglected philosophical questions. It contains many arguments and insights of publishable quality. In my view it easily meets the standards required of doctoral level work in philosophy of cognitive science and I am happy to **recommend the thesis for defense.**

There are a few ways in which I thought the dissertation could have been improved further still, and (unsurprisingly, given the sheer amount of interesting material that is surveyed) several points at which I'd be interested to know how the author wishes to further clarify or develop points raised in the dissertation. I have been asked to comment on the strengths of the dissertation as well as the areas where it could be improved. But in the interests of space I will not say much more about the dissertation's considerable strengths. Let me only repeat that the breadth and depth of knowledge on display is immensely impressive – all the more so given that the author is working in a relatively underdeveloped area of philosophy, and is often working with material that is very technical, very recent, or both.

There are three general points where I would be particularly interested in getting a clearer picture of the author's views.

- The first is largely a point about the way in which the dissertation as a whole is written and structured, rather than pertaining to any particular argument. The most impressive contribution of the dissertation is to bring such a great amount of technical material together, and work towards a unified conceptual framework in which it can all be placed to further our thinking about the unity of consciousness. The most important way in which I thought the dissertation could have been improved would have been to have the construction of this framework occupying centre stage far more frequently in the thesis. The conclusion does a good job of explaining, in broad terms, how the different parts of the thesis are intended to hang together. But at many points the reader would benefit from being reminded just where they are in the overall dialectic of the dissertation. For example, the discussions in chapters 2, 3, 4 and 5 combine to provide an informed view of just what the unity of consciousness is, the distinctive properties of unity which must be explained, and some of the constraints on theories which might seek to explain them. But there is no point in the dissertation where the lessons of these chapters are clearly brought together. It would have been particularly useful to spell out the lessons of these chapters heading into chapter 6, so that the criteria against which candidate cognitive scientific accounts of unity were being assessed were as clear as possible. If thinking about how to publish material from the thesis, or construct a monograph based on the content

of the thesis, doing more to bring out the single narrative thread which unites the elements of the thesis is the single biggest way in which I thought it could have been improved.

- The second point is more specific, and first arose during the discussion in chapter 3 – though I think it can be asked of the explanatory framework promoted by the thesis as a whole. The question is: Why should we think that practical unity (e.g. an agent manifesting behaviour that invites explanation in terms of the unity of some tokened representational states) is either necessary or sufficient for conscious unity? We might question both its necessity and sufficiency by holding (or perhaps just raising the possibility that) the conditions for *ascribing* unified consciousness come apart from the conditions for *possessing* unified consciousness. We might question its sufficiency by arguing that there could be cases where disunified bits of information processing nonetheless interact to produce adaptive, goal-directed behaviour – think for example of the way Rodney Brooks's subsumption architecture robots produce coherent, adaptive behaviour despite the fact that the information processed by their constituent modules is never brought together anywhere in the system. Are the materials set out in chapter 6 (and brought together in the conclusion) intended to address these questions?
- The final question on which I'd like to hear more from the author concerns the role of cultural and linguistic scaffolding in explaining the unity of consciousness. Hurley's work, I think, makes an implicit suggestion about this – the normative unity that is partly constitutive of unified consciousness is exhibited by us in virtue of our participation in socially scaffolded practices that have analogous normative structures. One such socially scaffolded practice is the linguistically mediated activity of 'giving and asking for reasons'. There were several places where I was unsure whether the author was endorsing a version of this thought. For example, in the discussion of GW theory on pp.113-4 it is suggested that linguistic scaffolding can solve a problem of how elements of a GW can 'talk to' each other, and a related suggestion appears to be made on p.131 for PP architectures. I wondered how the *public* shared code of language was here supposed to solve what appeared to be a problem about the need for a *neural* code that could be shared between domains or modules. More generally, I wondered about the precise role of socio-linguistic scaffolding was in supporting the cognitive and metacognitive capacities which the author argues are distinctive of unity.

As is to be expected in such a rich dissertation, there were many small points where questions arose, or I would have liked to hear a bit more. I list these below, in the order in which they occur in the thesis:

**Chapter 1:**

- Here I thought that the dialectical role of Dennett's work could be clarified – the level of detail/motivation wouldn't be enough to convince someone hostile to DD's ideas of the way they are used to frame the debate here. But perhaps DD's work is serving a more modest purpose here, of illustrating how the general set of problems with which the thesis is concerned can arise within an attempted naturalistic view of consciousness that purports to be aware of the danger of the homuncular fallacy?

**Chapter 3:**

- 3.1 (p.29): Why should it be the case that a naturalistic explanation of consciousness requires it to be a biological adaptation? Aren't there lots of phenomena that can be explained in a way continuous with the natural sciences (including evolutionary biology) that aren't biological adaptations?
- On p.38 it's claimed that neo-pragmatist accounts eschew neural representations – but wouldn't a more natural position for the neo-pragmatist be to hold that neural representations may be usefully and legitimately appealed to in certain discursive contexts, but that these contexts are importantly different from those in which we talk about personal-level representations and their contents?

- In this chapter (pp.36-7) and elsewhere in the thesis (especially chapter 6), I wondered about the precise sense of 'modularity' that was at issue in the discussion. Here the author claims that some notion of modularity is required to make sense of the notion of neural representation, and I'd welcome clarification of why this is. Later, in chapter 6, it's stated that:

  > 'A modular system would thus be limited to reflexive or habitual responses which do not take into account the situational context and are executed even if they are not relevant.'

  Without further clarification, it looks like this can't be right – think of language, the paradigmatic (for Fodor and others) candidate for a modular cognitive process. Our linguistic abilities, on their face, don't seem reflexive, habitual, context-insensitive or executed regardless of their relevance. So here and elsewhere I was puzzled about just what kind of modularity was at issue. It wasn't obvious to me what, if anything, in the rest of the thesis hinged on this particular understanding of modularity. I'd be interested to know if the author thought that a 'softer' understanding of modularity (e.g. as espoused by Wheeler and Clark in their paper 'Culture, Embodiment and Genes: Unravelling the Triple-Helix' would do the same dialectical work in the thesis.

**Chapter 4:**

- A minor question: do we really have reason to think that the dorsal pathway issues integrated representations of relative positions of objects and their movement? Don't the dual visual systems results suggest that the information carried by the dorsal stream is sparser than this, concerning only information suited for the fine control of visuomotor skills like grip scaling and orientation?

**Chapter 5:**

- On p.88-9, I wondered why Hurley's appeal to intentional access doesn't constitute a positive theory of how conscious contents are integrated. It's certainly not a detailed neuroscientific or computational theory – but doesn't it nonetheless do the job of specifying the functional relationships which Hurley thinks constitute perspectival self-consciousness?
- I also had some questions about the short evolutionary story about self-consciousness in 5.4.3:
  - Isn't the criterion of consciousness given on pp.91-2 *very* minimal such that various moderately sophisticated neural networks can meet it? The criterion appeared counterintuitively inclusive to me, and I'd welcome clarification about whether and why this is correct.
  - Is an *implicit* distinction between the subjective and objective order of things really enough to meet Strawson's requirement for transcendental self-consciousness? 'Implicit' suggests that this distinction is somehow *reflected in* the operations of the system, without being *represented by* the system. But, as with the above point, can't lots of simple learning systems in some way *reflect* the distinction between the world and their representation of it without plausibly being seats of transcendental self-consciousness? (p.92)
  - In the next paragraph – why is conceptualising a perceptual classification as a representation necessary for realizing that a perceptual error has been made? Isn't simply responding appropriately to the representation enough? For example, if a perceiver does a double-take because they seem to have perceived something unexpected, isn't this enough to indicate that they in some sense realise a perceptual error has been made – regardless of the terms in which they do or don't conceptualise that error?

**Chapter 6:**

- p.136: Isn't it intuitive that there is variability in the reliability of self-monitoring? Aren't being extremely tired, hungry, or under the influence of drugs are all examples of situations that give us good reason to be sceptical of self-reflective judgments?

- pp.146-7: I wondered how the proposal about relevance linked up with issues about the unity of consciousness here. How does the *relevance for adaptive behaviour* that IIT appeals to here link up with the *experienced relevance or salience* that seems to be at issue when exercising attention?
- p.149: How does the discussion of metarepresentation in IIT views here link up with the Kant/Hurley reading of the necessity of a metarepresentational *disposition* (which need not be exercised) for the unity of consciousness?
- p.155: I wasn't sure I was convinced by the putative example of partial unity here. Isn't the stated case one where the elements of the situation are part of a single unified experience (they are all co-conscious), but fail to be integrated in a way that guides optimal action? There's certainly a failure of optimal integration here, but why think that this constitutes a lack of fully unified consciousness?

Let me finish by saying again that I thought this was an extremely impressive dissertation. While it raised plenty of questions for me, this should be seen as a testament to the sheer scope and depth of the material it encompasses. I learned a lot from reading it, and very much look forward to discussing some of these issues with the author in future. To reiterate, I am happy to **recommend the thesis for defense.**

Sincerely,

Dr. Dave Ward
Lecturer in Philosophy