Something Mental Is Just in the Head, and What the Mental Out of the Head is Like

Abstract: In, "Why Nothing Mental is Just in The Head," Justin Fisher (*Noûs*, 2007) uses a novel thought-experiment to argue that every form of mental internalism is false. This paper shows that Fisher fails to refute mental internalism, and that a new variant of his example actually (a) confirms a form of mental internalism, as well as (b) John Locke's "resemblance thesis," thereby (c) disconfirming all externalist theories of mental content (the type of theory Fisher takes his original example to prove).

Key Words: intentionality, mental, content, internalism, externalism.

In, "Why Nothing Mental is Just in The Head,"[1] Justin Fisher uses a novel thought-experiment to argue that every form of mental internalism is false (where mental internalism is understood to be the view that, "an individual's mental features at a given time supervene upon what is in that individual's head at that time"[2]). This paper shows that Fisher fails to refute mental internalism, and that a new variant of his example actually (a) confirms a form of mental internalism, as well as (b) John Locke's "resemblance thesis,"[3] thereby (c) disconfirming all externalist theories of mental content (the type of theory Fisher takes his original example to prove).[4]

§1 reviews Fisher's thought-experiment and argument against mental internalism. §2 shows that Fisher mischaracterizes mental internalism, and that his example poses no threat to narrow functionalism, the view that some mental features supervene on the

[1] Fisher (2007).

[2] Ibid: 318.

[3] Locke's (1689: 2.8.15) thesis is, "the *Ideas of Primary Qualities* of Bodies, are *Resemblances* of them, and their Patterns really do exist in the Bodies themselves."

[4] See e.g., Dretske (1980, 1981, 1988, 1995), Millikan (1984, 1990, 1991, 1993), Papineau (1984, 1987, 1990, 1993), Stampe (1977), and Fisher (2007).

narrow functional roles internal states play “in the head.”[5] §3 constructs and defends a new variant of Fisher’s example. §4 shows that the example confirms narrow functionalism. §5 shows that the example confirms Locke’s resemblance thesis, thereby refuting externalist theories of mental content. Finally, §6 raises and responds to an externalist objection.

§1. Fisher’s Argument Against Mental Internalism

Fisher asks us to imagine two beings: a human being on Earth named ‘Edna’ and an alien on “Pulse World” named ‘Paula.’ Edna and Paula are supposed to be engaged in very different activities on their respective planets at particular time, *t*. Edna is supposed to be playing a saxophone on Earth (at t), whereas Paula is supposed to be driving a car along a Pulse World highway (at *t*). Finally, and most surprisingly, Edna and Paula are supposed to be in *precisely* the same internal state, *M*, at *t*. The explanation of how they are in the same internal state is this: Pulse World’s star emits a form of radiation that systematically interferes with neural processing. Consequently, Pulselings and humans evolved quite differently. Humans evolved such that *M* is implicated in saxophone-playing. Pulselings evolved such that *M* is implicated in car-driving.

Fisher expects some readers to be skeptical.[6] Could a single internal state really function so differently in different physical environments? Fisher goes to great lengths to demonstrate that the example is indeed metaphysically possible. Although some readers may have lingering doubts, the present paper will buttress Fisher’s argument. A clear

[5] Narrow functionalism is the view that some mental features supervene on the “narrow” functional roles of internal states, the functional roles those states play “purely in the head,” abstracting away from the external world.

[6] See Fisher (2007): §3.

proof of Fisher’s possibility (and a dramatic new possibility) will be provided in §3. For now, let us investigate Fisher’s analysis of the example.

Fisher claims that past externalist thought experiments (e.g. Putnam’s Twin Earth example and Davidson’s Swampman example) disprove only some forms of mental internalism.[7] For the sake of background, let us look at Fisher’s analysis of Putnam’s and Davidson’s examples, and then turn to his analysis of his own example. Putnam has us imagine two individuals: a person on Earth (“Earthling”) and a perfect replica of that person on “Twin Earth” (“Twin Earthling”). Besides obvious differences in physical location and numerical identity, there is only one difference between Earth and Twin Earth: whereas the stuff Earthling calls “water” has the chemical composition *H2O*, the stuff that Twin Earthling calls “water” has a very different chemical composition, *XYZ*. These facts are widely thought to show that some mental features depend on the external world. For whereas it seems clear that Earthling perceives (and has beliefs about) *H20*, it seems just as clear that Twin Earthling perceives (and has beliefs about) *XYZ*.[8] At the same time, Putnam’s example seems to confirm the internalist hypothesis that some mental features *are* just in the head. For although they have perceptions and beliefs about different objects in their environment, it seems clear that Earthling and Twin Earthling share a number of perceptions, thoughts, and beliefs “purely in the head.” They both have perceptions, beliefs, and thoughts as of “watery stuff.” Just as the stuff in Earth’s lakes and rivers *look watery* to Earthling, the stuff in Twin Earth’s lakes and rivers *looks watery* to Twin Earthling. These mental features appear to be “just in the head.”

Now turn to Davidson’s example. Davidson has us imagine that a fully formed

[7] Putnam (1973), Davidson (1987).

[8] Burge (1979) and (1982).

human being – "Swampman" – arises fully formed out of a swamp. We are to suppose that Swampman is a perfect internal duplicate of an actual human being (that he has all of the same internal states as, say, you or me). Davidson's example appears to show, much like Putnam's example, that some mental features depend on an organism's causal-historical relationships to the external world. For although Swampman may share all of my internal states, it seems clear that I have many mental features he lacks. I *remember* my childhood. Swampman just *seems* to remember his (he had none). I have beliefs about cars. Swampman only appears to have such beliefs (he has never even seen a car). And so on. At the same time, though, just like Putnam's example, Davidson's example appears to confirm the internalist hypothesis that some mental features are just in the head. For clearly, Swampman and I share many mental features. We both have memory states *as of* childhood, memory states *as of* having gone to graduate school, and so on.

Fisher maintains that his example demonstrates for the first time that *no* mental features are just in the head. His argument is as follows: Edna (the Earthling) and Paula (the Pulseling) are in *exactly* the same internal state, *M*, at *t*. Yet, Edna and Paula share *no* mental features at *t*. All of Edna's perceptions, beliefs, etc., at *t* are about *saxophones*. All of Paula's mental features at *t*, on the other hand – all of her perceptions, beliefs, etc. – are about *cars and freeways*. Since they share *no* mental features at *t* but are *exactly* the same "in the head" at *t*, mental internalism is false. Two creatures can be the same in the head at a given time and share no mental features at all.

§2. Mental Internalism Misconstrued and Narrow Functionalism Untouched

Fisher assumes, once again, that mental internalism is the view that, "an individual's mental features at a given time supervene upon what is in that individual's head at that time." This cannot be correct, however. For consider *narrow functionalism*: the view that some mental features supervene solely on the narrow functional role of internal states, the functional role those states play purely "in a person's head," abstracting away from the external world entirely. Narrow functionalism is plainly an internalist view. It is also philosophically compelling. It explains, among other things, why Earthling and Twin Earthling (and me and my Swampman duplicate) share mental content. It says that Earthling and Twin Earthling share "watery" perceptions and thoughts because their brains are doing the same thing internally (i.e. instantiating "watery" narrow functional roles). Similarly, it says that Swampman and I (as duplicates) duplicate share memories *as of* childhood, beliefs *as of* cars, etc., because *our* brains are doing the same thing internally (i.e. instantiating the very same "childhood-ish" and "car-ish" narrow functional roles). Narrow functionalism does not, however, satisfy Fisher's definition of mental internalism. For the functional characteristics of a given state are *not* comprised by what is the case at a given time. A state's functional role is comprised by the relations the state bears to *other* states, many of which only exist at other times. As an illustration, consider again the idea that Earthling and Twin Earthling share "watery" perceptions and thoughts. What comprises the "watery" narrow functional role of the state they share? Something can be individuated as "watery" only by contrasting it against *non*-watery things – by contrasting it against solid things, gassy things, and so on (e.g. watery substances are less dense than solid things but more dense than gasses). A state plays a "watery" functional role in a person's head, then, just insofar as it relates to other internal states that person may have at other times ("solidy" states, "gassy" states, and so on).

Mental internalism cannot, therefore, be correctly defined as the view that, "an individual's mental features at a given time supervene upon what is in that individual's head at that time." Mental internalism ought to be defined instead as the thesis that that an individual's mental features at a given time supervene on what is in that individual's head *at different times*. But if this is the right definition, then Fisher has not refuted mental internalism. In order to refute narrow functionalism, Fisher would have to show that two creatures could be in exactly the same internal *states across time* and yet share no mental features (at any of those times). We will now see, however, that this sort of case is impossible. Any two creatures who share perfectly similar internal states across time will share many features, regardless of how different their external environments may be.

§3. Super Pulse World

Fisher writes that "a general moral" can be extracted from his example, namely:

> The normal functioning of all cognitive systems deeply depends on their getting appropriate support (or at least appropriate non-interference) from their surroundings. For any complex cognitive system, there are possible surroundings in which that *system* would effectively perform cognitive control tasks completely different from those it normally performs.[9]

We will now see that Fisher is right about this, but that he has not fully fleshed out the idea's actually implications.

Let us call "Super Pulse World" a world in which a race of intelligent beings –

[9] Fisher (2007): 324-5, emphasis added.

Super Pulselings – evolved, thanks to a profoundly different physics, to use *in general* the kinds of internal states that Earthlings use to intelligently navigate Earth's environment to intelligently navigate a profoundly different *Super Pulse World* environment. Let us assume, in other words, that whereas Earthlings use a series of internal states (i.e. states A, B, C, D, etc.) to play saxophones, Super Pulselings use that *same* series to drive cars; and so on, across a wide array of behaviors, such that Edna the Earthling and "Super Paula" the Super Pulseling instantiate *exactly* the same internal states over their entire lives while, externally speaking, behaving in profoundly different ways.

Is Super Pulse World a genuine metaphysical possibility? It is surprisingly easy to demonstrate that it is. Begin with the case of visual cognition. Two different types of facts – empirical facts about the human visual cortex and first-personal phenomenal introspection – both strongly support the idea that human visual processing is fundamentally "picture-like" (i.e. involving the construction of internal mental *images* – images that are in many respects similar to ordinary photographs). First, the human primary visual cortex has a *retinotopic* structure. It is a *neural map* of the person's visual field. Individual neurons in the primary visual cortex appear to function much like individual "pixels" in an ordinary bit-map visual display (with individual neurons representing *points in* visual space).[10] When a human being looks at a saxophone, her primary visual neural array fires in a *distinctly saxophone-shaped* configuration. The primary visual cortex appears, in other words, to create an "internal mental picture" of the person's visual field. This idea, of course, fits well with instrospection. We not only use

[10] See Dougherty et al (2003), Rosa (2003), and Wandell et al (2005).

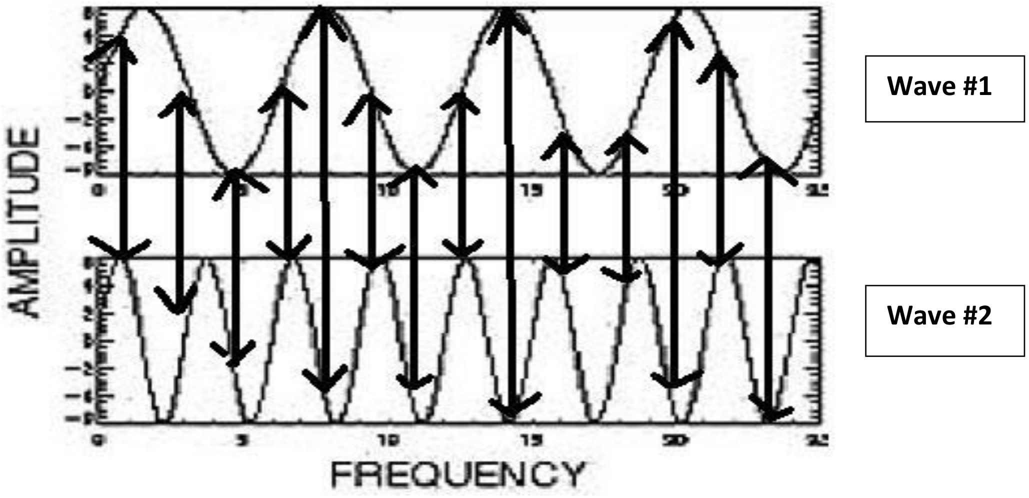
the language of pictures to describe our visual experiences (we say things like, "I have an *image* of Jane's face in my mind right now"); we also describe our visual experiences in much the same way that we describe ordinary pictures (e.g., we say things like, "*This* point in my visual field is blue, *this* point is red, etc.").

We can now show that Super Pulse World is at least visually metaphysically possible (i.e. Possible with respect to visual cognition). For it can be shown that any picture of one type of object (e.g. a picture of a saxophone) can *function* as, or serve as, a picture of a very different object (e.g. a road). Here is the demonstration. Consider an ordinary picture of a saxophone (as projected on a computer screen). This picture is nothing more than a series of pixels of different colors (assume for the sake of argument that "Pixel #1" has the value "gold", that "Pixel #2" has the value "silver," and so on). Now turn to a similar picture of a road (on the same computer screen). This picture too is nothing more than a series of pixels of different colors (assume "Pixel #1" has the value "black," "Pixel #2" has the value "grey," and so on). Here then is the question: can the picture of the saxophone be used to (or function to) represent *everything* that the picture of the road represents (namely, the surface features of a road)? Clearly. All one needs to do is to apply the following translation rule to the saxophone image: "Change Pixel #1 from black to gold, change Pixel #2 from grey to silver, etc." This rule will for all intents and purposes turn the saxophone picture *into* the road picture. Finally – and this is the critical part – this translation rule could at least in principle be applied *externally* to the picture. So, for example, suppose the picture of the saxophone is literally in a person's head (i.e. coded into the retinotopic map in the primary visual cortex). Here is one way mapped onto an external reality of roads: simply build the translation rule (mapping the saxophone picture onto a road picture) into the physics of the world *outside* of the person’s head. By building the translation rule into the physical laws of the world outside of the person’s head – building the rule into physical laws *relating* internal states to the external world – the picture of the saxophone in the person’s head could *function* as a picture of a road, enabling intelligent road-driving behavior. Notice, indeed, that this is true of any *series* of pictures. Insofar as every picture in a *series* of saxophone pictures could be mapped onto a corresponding picture in a series of road pictures, any series of saxophone pictures (“in a person’s head”) could, at least in principle, be used to represent the surface features of roads (outside of the head, in the external world).

Now obviously, building such translation rules into the physics of a world would be an incredibly complex and gerrymandered affair. Consider, after all, what one would have to *do* in order to map a series of saxophone pictures onto a series of road pictures. One would have to painstakingly map *every* value of *every* pixel of *every* saxophone picture onto corresponding pixels of the road photographs. Super Pulse World is, therefore, highly improbable. It is incredibly unlikely that such a physics exists anywhere in our Universe. The mere fact that a physics is (nomologically) unlikely, however, is *no* reason to think that it is metaphysically impossible. In order to show that something is metaphysically possible (at least according to common philosophical practice), all we have to do is show that it is *conceivable*. And we have shown that Super Pulse World’s physics is conceivable. Insofar as there is always some possible mapping from a saxophone photograph to a road photograph, it is metaphysically possible in principle for a series of saxophone-y visual states “in the head” to function as representations of an external world of roads. A god (or Cartesian Demon) could in principle create such a physics.[11] Super Pulse World is, therefore, metaphysically possible (at least in the case of visual representation).

This argument can be extended straightforwardly to other types of sense perception. For example, consider auditory representation. Insofar as different sound waves differ only in terms of frequency and amplitude, there is always in principle a possible mapping rule to "translate" one sound wave into another wave. So, for example, consider the two sound waves in Figure A. Wave #1 can be translated into Wave #2 by simply mapping every point on Wave #1 to every point on Wave #2 (as indicated by the arrows in Figure I). The fact that such a translation exists, however, is just to say that any cognitive representation of Wave #1 can be *transformed* into a cognitive representation of Wave #2 (by simply applying the translation rule).

Figure I



Now, again, mapping sound waves to one another in this sort of way is a highly gerrymandered affair. Still, the point is that it is *metaphysically possible*. Fisher, then,

[11] For an influential defense of this view, see Chalmers (2002). Readers interested in the philosophical debate about the relationship of conceivability to possibility should also see Gendler and Hawthorne (2002).

was right when he wrote, “For any complex cognitive system, there are possible surroundings in which that system would effectively perform cognitive control tasks completely different from those it normally performs.”[12] Cognitive systems *can* in principle be mapped onto, and so produce very different types of intelligent behavior, in profoundly different environments.

§4. Narrow Functionalism Confirmed

What, if anything, does the Super Pulseling example show? It clearly confirms narrow functionalism. For although Edna and Super Paula’s *bodies* are doing very different things (e.g., Edna’s body is playing a saxophone from t-t_n whereas Super Paula’s body is driving a car along a Pulse World highway during t-t_n), it is clear that they have the same internal mental picture: a *qualitatively saxophone-y* picture (one whose elements *qualitatively* correspond to the surface features of saxophones, not the surface features of roads). Edna and Super Paula share a number of internal mental features – their minds are saxophone-y “on the inside” – for the simple reason that their brains are, at least internally speaking, functioning in a *saxophone-y* way. Edna and Super Paula have the same *saxophone-y phenomenal experiences*, the same *saxophone-y phenomenal beliefs*, and so on.

§5. Locke’s Resemblance Thesis Confirmed and Pure Externalism Disconfirmed

What does Super Paula perceive, or believe, when it comes to the external world? Her body certainly drives her car in what looks to be a highly intelligent manner. But what are

[12] Fisher (2007): 324-5 (emphasis added).

Super Paula's mental contents, *really*? Does she *see* the road in front of her body? Does she have *beliefs* about it? Although it may be tempting to say that she does, thanks to her body's behavior, logical behaviorism is widely agreed false. Behavior does not a mental state make. In order to know what a person perceives or believes, we should look carefully at the biological organ we know to *do* perceiving, believing, and other mental tasks: the brain. With this in mind, let us look at what Super Paula's brain does.

Consider Super Paula's brain. What does her brain indicate about *her*? Super Paula's body intelligently navigates a world of roads. Her brain does not, however, *present the world to her, qualitatively, as the world actually is*. When her body navigates roads, Super Paula's brain instantiates *qualitatively saxophone-y* internal mental pictures. How, then, can Super Paula be said to *see*, *perceive,* or have *beliefs* about the world of roads that stretch out in front of her body? Her *conscious* mental life – the realm of her conscious experience – simply does *not* present her with a world of roads. Super Paula is mentally cut off from the external world in the very same way that the classical brain in the vat is cut off from its (virtual-reality) world. She does not *see* the roads in her environment any more than the brain in the vat *sees* the computer code it is being fed (computer code giving it the *illusion* of seeing tables, chairs, and other people). Upon reflection, it is only Edna (the Earthling) who really sees the external world. Edna *sees* saxophones because, in addition to bearing the right kinds of causal/external relations to saxophones, Edna has something that Super Paula lacks: internal representations that *qualitatively resemble* saxophones.

The Super Pulse World case thus demonstrates that Locke's resemblance thesis is correct and purely externalist theories of mental content are false. Super Paula instantiates *every* relevant externalist relation to objects in her external environment, yet she lacks genuine perceptions or beliefs about its features. In order to perceive and have other mental contents (e.g. beliefs) about the external world, a being's internal cognitive states must (at least in general) *qualitatively* represent the world as it really is.[13]

§6. An Externalist Rebuttal?

Externalists might be tempted to reply as follows: "Even though her experiences are qualitatively unlike the external world, there is a very clear sense in which all of Super Paula's mental states – her perceptions, beliefs, etc. – are *about* features in her external environment. After all, her mental states *detect* features of her world (e.g. roads). Insofar as her states detect these features, they are *about* them (intentionally, and so mentally, speaking)."[14] All of this is right. There is *a* sense in which Super Paula's mental states are about the external world. Her mental states *do* detect features of her environment. The problem, however, is with the very sense in which they detect those features. Super Paula's mental states detect features of her external environment only in a *behavioral* sense – in the sense that they (the mental states) are used *by her body* to navigate roads. The problem is that this is behaviorism. Few, if any, philosophers today believe that bodily behavior itself is "the mark of the mental." Behavior is just that: behavior. *Consciousness* is the mark of the mental. The Super Pulseling case demonstrates that we

[13] It is worth noting that the Super Pulseling case is not just an interesting new variant of the classical brain-in-the-vat hypothesis (Super Paula is "envatted" in the sense that her brain does not present the world to her as it really is), but that it is also a new variant of John Searle's (1980) "Chinese Room." For just as the person in Searle's room uses a series of rules to produce fluent Chinese (without actually *knowing* any Chinese), Super Pulse World's "translation rules" enable Super Paula's *body* to produce intelligent road-driving behaviors even though her *mind* has no idea, internally, that her body is driving cars (again, the world *looks* saxophone-y to her).

[14] Fisher has proposed this reply in personal communication.

only truly see the world (in a robustly *mental*) sense – I only *really* have beliefs about saxophones – to the extent that the world *phenomenally (or consciously)* appears to us as it is. The Super Pulseling case demonstrates, in other words, that there are two very different kinds of intentionality: purely behavioral intentionality, and genuinely *mental* intentionality. Externalist theories of "mental content" may be adequate accounts of the former; they are not, however, adequate accounts of the latter.

References

Burge, Tyler (1979). “Individualism and the Mental,” in *Studies in Metaphysics*. P. French, T. Uehling, and H. Wettstein, eds., Minneapolis: University of Minnesota Press.

----- (1982). “Other Bodies,” in Andrew Woodfield, ed., *Thought and Object*. New York: Oxford.

Chalmers, David (2002). “Does Conceivability Entail Possibility?”, published in T. Gendler and J. Hawthorne, eds., *Conceivability and Possibility*, (Oxford: Oxford University Press, 2002): 145-200.

Davidson, Donald (1987). “Knowing One’s Own Mind,” *Proceedings and Addresses of the American Philosophical Association*, 60: 441-58.

Dougherty, R. F., Koch, V. M., Brewer, A. A., Fischer, B., Modersitzki, J., & Wandell, B. A. (2003). Visual field representations and locations of visual areas V1/2/3 in human visual cortex. *Journal of Vision*, 3(10):1, 586-598, http://journalofvision.org/3/10/1/, doi:10.1167/3.10.1.

Dretske, Fred (1980) “The intentionality of cognitive states,” in Rosenthal, D. (ed.)(1990) *The Nature of Mind*, Oxford: Oxford University Press.

_______ (1981) *Knowledge and the Flow of Information*, Cambridge, Mass.: MIT Press.

_______ (1988) *Explaining Behavior*, Cambridge, Mass.: MIT Press.

______ (1995) *Naturalizing the Mind*, Cambridge, Mass.: MIT Press.

Gendler, Tamar and Hawthorne, John. (2002). *Conceivability and Possibility*, (Oxford: Oxford University Press).

Fisher, Justin (2007). "Why Nothing Mental Is Just in the Head," *Nous*, 41 (2): 318-334.

Locke, John (1689). *An Essay Concerning Human Understanding*.

Locke, John (1975). *An Essay Concerning Human Understanding*, ed. P. Nidditch, Oxford: Clarendon Press (original work 2nd ed., first published 1694)

Millikan , Ruth Garrett. (1984) *Language, Thought and Other Biological Objects*, Cambridge, Mass.: MIT Press

______(1990). "Truth, Rules, Hoverflies and the Kripke-Wittgenstein Paradox" in *Philosophical Review*, 99: 232-53.

______(1991). "Speaking Up for Darwin" in Loewer, B. & Rey, G. (eds.) (1991) *Meaning in Mind: Fodor and his critics*, Cambridge, MA: Blackwell, 151-165.

_______(1993). *White Queen Psychology and Other Essays for Alice*, Cambridge, MA: MIT Press.

Papineau, David (1984). "Representation and Explanation", in *Philosophy of Science*, 51: 550-72.

_______(1987). *Reality and Representation*, Oxford, UK: Basil Blackwell.

_______(1990). "Truth and Teleology" in D. Knowles (ed) *Explanation and its Limits* Cambridge University Press, 21-44.

_______(1993). *Philosophical Naturalism* Blackwell, Oxford.

Putnam, Hilary (1973). “Meaning and Reference,” in *The Philosophy of Language*. A.P. Martinich, ed., New York: Oxford University Press, 1996.

Rosa, MGP (2002). “Visual maps in the adult primate cerebral cortex: some implications for brain development and evolution.” *Brazilian Journal Medical and Biological Research*, 35(12):1485-1498.

Searle, John (1980). “Mains, Brains, and Programs,” *Behavioral and Brain Sciences*, 3 (3): 417-457.

Stampe, D., (1977). "Toward a Causal Theory of Linguistic Representation", in P. A. French, T. E. Uehling, Jr., and H. K. Wettstein (eds) *Midwest Studies in Philosophy: Studies in the Philosophy of Language*, vol. 2, Minneapolis: University of Minnesota Press, 81-102.

Wandell, B., A. A. Brewer and R.F. Dougherty (2005). “Visual Field Map Clusters in Human Cortex,” *Philosophical Transactions of the Royal Society B*, vol. 360, pp. 693-707.