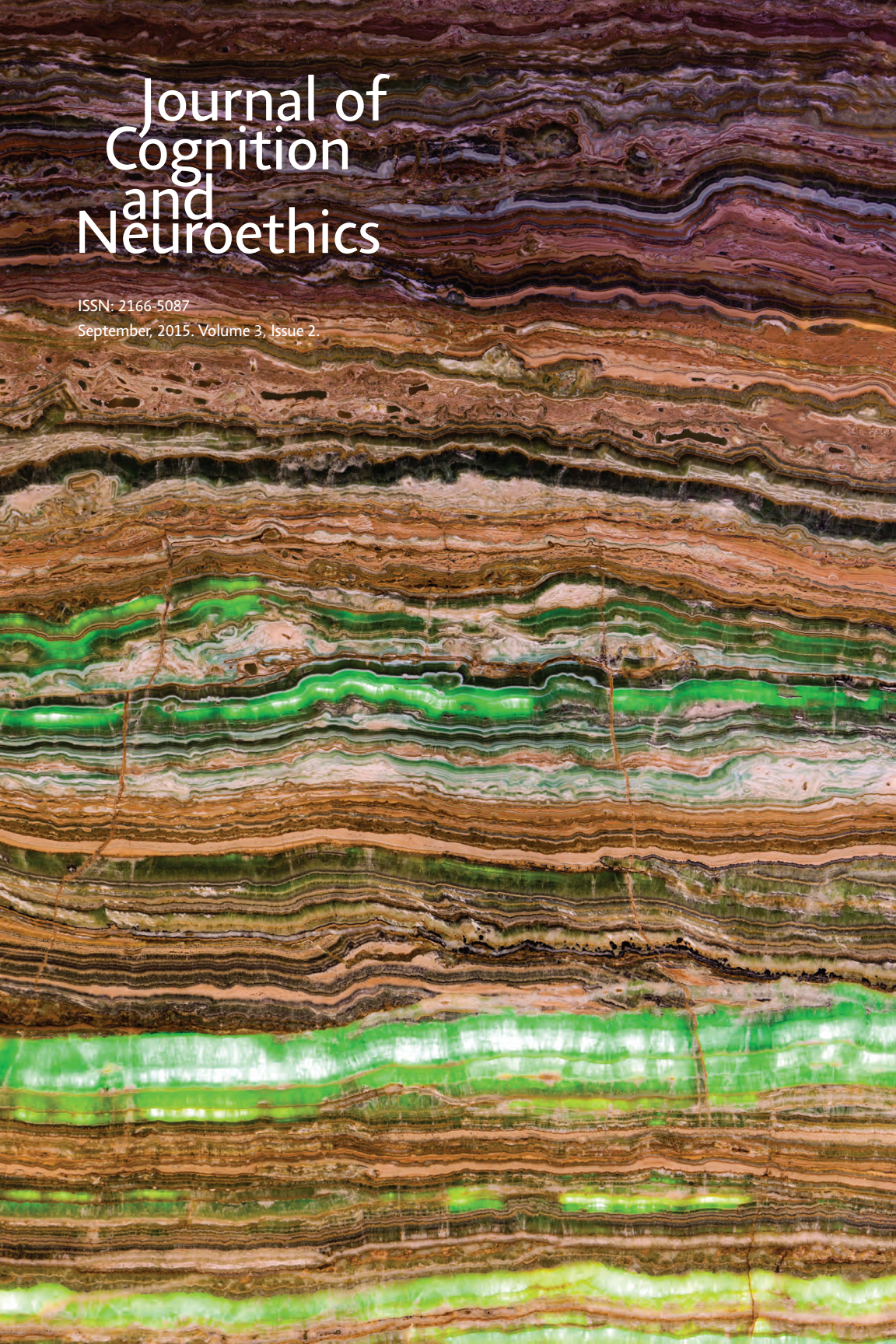

ISSN: 2166-5087
September, 2015. Volume 3, Issue 2.

Journal of Cognition and Neuroethics

Managing Editor
Jami L. Anderson

Production Editor
Zea Miller

Publication Details
Volume 3, Issue 2 was digitally published in September of 2015 from Flint, Michigan, under ISSN 2166-5087.

The *Journal of Cognition and Neuroethics* is produced by the Center for Cognition and Neuroethics. For more on CCN or this journal, please visit cognethic.org.

Center for Cognition and Neuroethics
University of Michigan-Flint
Philosophy Department
544 French Hall
303 East Kearsley Street
Flint, MI 48502-1950

Table of Contents

Journal of Cognition and Neuroethics

Rescuing PAP from Widerker's Brain-Malfunction Case

Greg Janzen
Osgoode Hall Law School

Biography

Greg Janzen has been a lecturer in philosophy at Mount Royal University and the University of Calgary. Some of his work has appeared in *Journal of Consciousness Studies*, *Philosophia*, *Journal of Philosophical Research*, and *Erkenntnis*. He is also the author of *The Reflexive Nature of Consciousness* (John Benjamins Publishing, 2008). He is currently attending Osgoode Hall Law School in Toronto.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). September, 2015. Volume 3, Issue 2.

Citation

Janzen, Greg. 2015. "Rescuing PAP from Widerker's Brain-Malfunction Case." *Journal of Cognition and Neuroethics* 3 (2): 1–22.

Rescuing PAP from Widerker's Brain-Malfunction Case

Greg Janzen

Abstract

According to the principle of alternate possibilities (PAP), a person is morally responsible for what she has done only if she could have done otherwise. David Widerker, a prominent and long-time defender of this principle against Harry Frankfurt's famous attack on it, has recently had an unexpected about-face: PAP, Widerker now contends, is (probably) false. His rejection of PAP is a result, in large part, of his coming to believe that there are conceptually possible scenarios, what he calls 'IRR-situations,' in which circumstances that nowise bring it about that an agent performs a particular action are precisely the circumstances that make it impossible for her to avoid performing that action. The circumstances that guarantee that the agent will perform the action turn out to be immaterial, since the agent in an IRR-situation is blameworthy because she would have performed the action even if, contrary to the specified facts, an alternative course of action had been available to her. The goal of this article is to show that Widerker's report of PAP's demise has been an exaggeration: careful scrutiny reveals that the kind of scenario that he believes refutes the principle—his 'brain-malfunction' scenario—is not an IRR-situation at all.

Keywords

Principle of Alternate Possibilities, Free Will, Moral Responsibility, Harry Frankfurt, Frankfurt Cases, Compatibilism, David Widerker

Introduction

According to the principle of alternate possibilities (PAP), a person is morally responsible for what she has done only if she could have done otherwise.[1] This principle

1. So stated, PAP is equivalent to Frankfurt's original formulation, which says that 'a person is morally responsible for what he has done only if he could have done otherwise' (1969, 829). But this formulation admits of at least two interpretations: (i) as expressing the proposition that a person is morally responsible for performing a particular action, *V*, only if she could have *avoided* (or *refrained* from) doing *V*, or (ii) as expressing the proposition that a person is morally responsible for *V* only if she could have avoided doing *V* and, in addition, could have performed *some other action*. This latter interpretation is very non-standard—neither friends nor critics of PAP interpret it thus—and there is no reason to interpret it thus. Accordingly, PAP, in what follows, should be taken as expressing the proposition that an agent is morally responsible for doing *V* only if he could have avoided doing *V*. A further bookkeeping point is that, unless otherwise noted, 'moral responsibility' should be regarded as synonymous with 'morally blameworthy,' since Widerker's project (in his 2006) is to develop an account of moral blameworthiness. Compare his

is indisputably philosophically important, frequently playing a pivotal role in debates regarding the sort of freedom required for moral responsibility. David Widerker, a prominent, eloquent, and long-time defender of the principle against Harry Frankfurt's (1969) famous attack on it (Widerker 1995; 2000; 2003; 2005), has recently had an unexpected about-face. PAP, Widerker now contends, is (probably) false: an agent can be morally responsible for what she has done even though she could not have done otherwise.[2] He now subscribes to a view that he calls 'Frankfurt-friendly libertarianism,' which is a libertarian-based account of freedom and moral blameworthiness that rejects PAP (Widerker 2006, 2009). Widerker's rejection of PAP is a consequence, in large part, of his coming to believe that there are conceptually possible scenarios, what he calls 'IRR-scenarios' or 'IRR-situations,' in which circumstances that nowise bring it about that an agent performs a particular action are precisely the circumstances that make it impossible for her to avoid performing it. The circumstances that guarantee that the agent will perform the action turn out to be immaterial, since the agent in an IRR-situation is blameworthy because she would have performed the action even if, contrary to the specified facts, an alternative course of action had been available to her.[3] The goal of this article is to show that Widerker's report of PAP's demise has been an exaggeration: careful scrutiny reveals that the kind of scenario he believes refutes the principle is not an IRR-situation at all.

1. Widerker's rejection of PAP

Let me begin by explaining how the axis of Widerker's examination regarding PAP became rotated so that he now rejects the principle.

Widerker (2006, 168) characterizes an IRR-situation as follows (where a state of affairs is *causally possible* at time *t* if and only if it obtains in some possible world that up until, and including, *t* has the same history and the same laws of nature as the actual

rendering of PAP: 'An agent *S* is morally blameworthy for performing a given act *V* only if he could have avoided performing it' (2006, 163).

2. This is not to say that Widerker now believes that PAP lacks intuitive appeal. On the contrary, he still considers it very intuitively appealing, since it is accords with 'the intuitive link between attributing blame to an agent and expecting of the agent that he not have done what he did' (Widerker 2009, 90).

3. It is not clear to me why Widerker calls them IRR-situations, but my guess is that it is because the circumstances that close off alternative possibilities are *irrelevant* to whether the agent performs the action in question.

world, and where an alternative is *actionally accessible* to an agent at *t* if and only if the agent has it within her power to bring about that alternative at *t*):

a. At t, S decides-to-V on his own.
b. *P1, P2,…, Pn* are *all* (in the circumstances) causally possible alternatives to S's deciding-to-V at t.
c. Each Pi (i = 1 – n) is *actionally inaccessible* to S in the circumstances.

This characterization of an IRR-situation, Widerker claims, is in keeping with Frankfurt's (1969, 830, 837) original characterization as a situation in which the agent's decision, though unavoidable, is in no way brought about by the circumstances in which it occurs.[4] Widerker now believes that IRR-situations are conceptually possible, and he has developed an example of such a situation himself.

What is the significance of IRR-situations, assuming that they are conceptually possible? Do they imply that PAP is false? In the past, Widerker was reluctant to accept this conclusion, because if PAP is false, one has to assume that an agent in an IRR-situation who knowingly performed a wrong action is blameworthy despite being unable to avoid performing it. But Widerker no longer finds this assumption objectionable. 'I find it intuitive,' he remarks, 'that in [certain IRR-situations] the agent is blameworthy for the decision he made, even though it was not within his power not to make it' (2009, 91). By Widerker's lights, then, PAP has been falsified, since the following two conditions for falsifying the principle have been satisfied: (1) an IRR-situation has been described, and (2) the agent in the IRR-situation is blameworthy for his action.

By way of introducing Widerker's example of an IRR-situation, consider the following purported Frankfurt-style counterexample to PAP that, on his view, does *not* describe an IRR-situation:

4. Widerker also provides a more general notion of an IRR-situation, which he calls a 'G-IRR-situation'. In a G-IRR-situation, the agent lacks a *morally significant alternative*, where *V** is a morally significant alternative to a morally wrong act *V* that an agent *S* performs at *t* if and only if: '(i) *V* and *V** are incompatible, (ii) *S* is aware (or should be aware) that if he performs *V** at *t*, he will not be acting in a morally wrong way, (iii) *S* truly believes that it is within his power to perform *V** at *t*, and (iv) in the circumstances, *V**-ing at *t* would be regarded as a reasonable way for *S* to avoid *V*-ing at *t*' (2006, 175). As Widerker points out, since the notion of a G-IRR-situation is more general than that of an IRR-situation, constructing an example of a G-IRR-situation should be easier than constructing an example of an IRR-situation. For the time being, I will disregard the distinction between IRR-situations and G-IRR-situations and continue to speak only about the former. The notion of a morally significant alternative will remerge, however, in my argument (in the next section) that Widerker has not constructed an IRR-situation.

> PROMISE-BREAKING (PB): Jones is deliberating as to whether to keep the promise he made to his uncle to visit him in the hospital shortly before a critical operation his uncle is about to undergo. Jones is his uncle's only relative, and the visit is very important to the uncle. The reason for Jones's deliberating is that, on his way to the hospital, he (incidentally) met Mary—a woman with whom he was romantically involved in his distant past and whom he has not seen since then. Mary, being eager to talk to Jones, invites him for a cup of coffee in a nearby restaurant. She explains that she is in town just for a couple of hours, and wishes to spend those hours with him. Jones is aware that if he accepts Mary's offer, he will not be able to make it to the hospital during visiting hours. Unbeknownst to Jones, there is another person, Black, who for some reason does not want Jones to visit his uncle. Black has the power and the means to force Jones to decide to stay with Mary. But wishing to avoid showing his hand unnecessarily, he has made up his mind to intervene if and only if Jones does not show a sign that he is going to decide to break his promise to his uncle. Call that sign Q. If Jones shows that sign, then Black does nothing, knowing that in this case Jones will decide to accept Mary's invitation. (It is assumed that Black knows Jones very well in this regard.) Finally, suppose that Black does not have to intervene, since Jones decides on his own not to keep the promise, so as to be able to spend time with Mary. (Widerker 2006, 164)

Here it would seem that Jones is blameworthy for breaking his promise because Black exerted no influence on his behaviour (even though he made it impossible for him to avoid breaking it).

According to Widerker (2006, 165–6), however, PB founders on the dilemma objection.[5] Exponents of this objection advance the following dilemma regarding the relation between Q, absence of which would have led Black to intervene, and Jones's decision. Either Q (or a condition whose presence is indicated by Q) determined Jones's decision, given the circumstances, or it did not. If it did, then the decision was brought

5. This objection, which has been suggested by a number of philosophers, most notably Widerker (1995), Ginet (1996), and Kane (1996, 142–4, 191–2), has spawned a small sub-industry in the literature on the debate over Frankfurt-style counterexamples. For important critical responses to it, see Mele and Robb (1998), Fischer (1999; 2010), McKenna (2003), and Haji and McKenna (2004).

about by the conditions in which it occurred and thus we do not have an IRR situation. If, on the other hand, Q was merely a reliable indicator of Jones's decision, then he might have exhibited Q, which would have prompted Black to refrain from intervening, but decided to keep his promise anyway. On the first horn of the dilemma, Jones's decision was determined to take place given the conditions in which it was made, and on the second horn, he could have decided otherwise. Either way, PB is not an IRR-situation.

Widerker, though, now believes that the dilemma objection fails,[6] for there are scenarios that, unlike PB, *do* describe an IRR-situation and that are not vulnerable to dilemma worries. The following variant of PB, he claims, describes such a situation:[7]

> BRAIN-MALFUNCTION (BMF): As in Promise Breaking, Jones deliberates as to whether to accept Mary's offer, and ultimately decides on his own at *t* to violate the promise to his uncle. Normally, one can avoid deciding as one does by deciding otherwise. But in our scenario Jones does not have that option, since shortly before beginning to deliberate, he undergoes a neurological change as a result of which one of the (neurological) causally necessary conditions for his deciding otherwise, a condition which we may call *N*, does not obtain. Let us also assume that all this is unknown to Jones (who believes that he can decide to keep the promise), and that *N*'s absence does not affect his deliberation process. Note that these assumptions do not render Jones's actual decision of not keeping the promise (D(– K)) causally determined. The fact that *N* is a causally necessary condition of Jones's deciding to keep the promise entails that *N*'s absence is sufficient for its not being the case that Jones decides to keep it, that is, for – D(K). But from this it does not follow that *N*'s absence is sufficient for D(– K)…. [W]e have not yet eliminated all the alternatives to Jones's deciding not to keep the promise that are accessible to him. There is the alternative of his continuing to deliberate at *t*, instead of making D(– K). However, this alternative can be ruled out either by stipulation or by assuming that in the scenario at hand Jones needs to make a

6. Though he still defends it against certain criticisms: see Widerker and Goetz (2013).

7. See McKenna (2003, 209–10) for a similar example, which, as it happens, he developed jointly with Widerker. My argument that Widerker's example is not an IRR-situation applies, *mutatis mutandis*, to McKenna's.

> decision right away, as otherwise he will miss the bus that can get him to the hospital on time. If that is the case, then in our scenario, the option of continuing to deliberate is practically equivalent to deciding not to keep the promise. (Widerker 2006, 169–70)

This scenario is not vitiated, it would seem, by the dilemma objection, since there is no sign Q that could generate the dilemma between lack of libertarian freedom, on the one hand, and the presence of an actionally accessible alternative, on the other. Thus, on Widerker's view, BMF satisfies the criteria for an IRR-situation: 'In it, the only alternatives to [Jones's deciding not to keep his promise] are alternatives which, though causally possible, are actionally inaccessible to Jones' (2006, 171).

But is BMF an IRR-situation? In the next section I will attempt to show that it is not, and hence that it is not deleterious to PAP. As we have seen, an IRR-situation must feature an agent who (1) is blameworthy for his decision and (2) lacks an actionally accessible alternative. I shall argue that BMF fails to meet this second requirement.[8]

2. BMF is not an IRR-situation

The reason that BMF is not an IRR-situation is that Jones has an actionally accessible alternative: instead of deciding to break his promise at *t*, he can continuing deliberating (in a way shortly to be elucidated).[9] Widerker, as we have seen, rejects this alternative on the ground that it is not actionally accessible. He says, to repeat, that:

> [the alternative of continuing to deliberate] can be ruled out either by stipulation or by assuming that in the scenario at hand Jones needs to make a decision right away, as otherwise he will miss the bus that can get him to the hospital on time. If that is the case, then in our

8. I am not the first person to defend PAP against Widerker's attack on the principle. At least one other commentator, Carlos Moya (2007), has also taken up the challenge, arguing that BMF fails to meet the first requirement. For a cogent reply to Moya, see Widerker (2009). (In section 2.2 I briefly discuss Moya's argument and Widerker's rejoinder.) Widerker (2006, 169) has proposed another example of an IRR-situation, which he calls '*Z-persons*,' but I will not discuss that example here. Although Moya's argument that BMF fails to meet the first requirement for an IRR-situation does not refute BMF, it does, it seems to me, refute *Z-persons*.

9. Note that since it is inevitable that Jones will fail to keep his promise after his neurology fails him and he begins deliberating, the claim being disputed is that he has no alternative to *deciding* to break it. This decision, I contend, is not one that he has to make.

> scenario, the option of continuing to deliberate is practically equivalent to deciding not to keep the promise. (2006, 170)

But there is ample room for disagreement here. My strategy will be to show that the alternative of continuing to deliberate cannot be foreclosed in the ways Widerker describes. I first, in section 2.1, argue that it cannot be ruled out by adding a time constraint. Then, in section 2.2, I argue that it cannot be ruled out by stipulation.

2.1 The alternative of continuing to deliberate cannot be ruled out by adding a time constraint

According to Widerker, once we assume that Jones needs to make a decision right away (as otherwise he will, for example, miss the bus that can get him to the hospital on time), 'the option of continuing to deliberate is practically equivalent to deciding not to keep the promise'. But is it?

There are two ways to interpret this equivalence claim: (i) as the claim that continuing to deliberate is *morally* equivalent to deciding to break the promise, or (ii) as the claim that continuing to deliberate is *one and the same* (mental) action as deciding to break the promise.[10] I will examine these interpretations in turn.

On this latter interpretation, the equivalence claim seems clearly false. If Jones continues deliberating—say because he is conflicted about what to do and is unable to make a decision—then he has not made a decision at all. Quite the reverse, he has been *indecisive*, a not uncommon phenomenon even among morally competent agents, and his indecisiveness has resulted in his breaking his promise. It is true that Jones *knows* (at some level) that if he does not make a decision by *t*, then, at *t*, he will break his promise, but it does not follow from this that if he fails to make a decision by *t*, then, at *t*, he has decided to break his promise. It is more accurate to say that Jones's breaking his promise is a *foreseeable consequence* of his continuing to deliberate, not that his continuing to deliberate is the same action as his deciding to break his promise.

It merits emphasizing that Widerker understands 'action' in Ginet's (1990) sense, i.e., as 'either a causally simple mental action such as a volition-to-*V* or the forming of

10. An anonymous reviewer has suggested a third interpretation of the equivalence claim: that continuing to deliberate is *tantamount* to deciding to break the promise, without being either morally or numerically equivalent to breaking it. If, however, the former is tantamount to the latter, then the former is equivalent in seriousness to the latter, so it would seem that on a 'tantamount' interpretation of the equivalence claim, continuing to deliberate is morally equivalent to deciding to break the promise. In short, the tantamount interpretation collapses into the moral equivalence interpretation.

an intention-to-*V*, or a complex action such as Sam's action of killing Smith that consists of a simple mental action causing the event of Smith's death' (Widerker 2006, 176n26). Notice that, in continuing to deliberate, Jones has neither a volition nor an intention to break his promise. Indeed, it is precisely the *lack* of those elements in his practical reasoning process that makes it the case that his continuing to deliberate is not the same action as his deciding to break his promise. Thus, on Widerker's own understanding of what constitutes an action, Jones's continuing to deliberate is not the same action as his deciding to break his promise.

The equivalence claim, then, is better understood as the claim that the act of continuing to deliberate is morally equivalent to deciding to break the promise. And, in fact, in his 2009 restatement of his 2006 argument for the conceptual possibility of IRR-situations, Widerker states the equivalence claim using moral equivalence terminology. '[T]he act of continuing to deliberate,' he says, 'is morally equivalent to the decision not to keep the promise, and hence does not count as a *morally significant alternative*' (2009, 90).[11] The idea is that, even if continuing to deliberate and deciding to break the promise are not the same action, continuing to deliberate is not an actionally accessible alternative, since even if Jones availed himself of this option, he would not avoid blame.

But this is eminently disputable. Following Widerker, I will call an alternative that is actionally accessible in the moral sense a 'morally significant alternative'. Despite initial appearances, Jones's continuing to deliberate is such an alternative. To see this, consider the following counterfactually modified version of BMF in which Jones undergoes a moral struggle as a result of the absence of *N*:

> BRAIN-MALFUNCTION* (BMF*): Jones, while deliberating in a normal fashion about whether to keep his promise, adopts a favourable attitude toward keeping it and, as a result, forms a desire to keep it. At this point, due to the absence of *N*, he hits a wall (so to speak). His neurological deficit prevents him from acting on his desire and deciding to keep it. His desire to keep it persists, however, and he verges on so deciding. At this point, *N*'s absence again prevents him from acting on his desire, so he continues deliberating. This back-and-forth goes on until the moment he runs out of time to deliberate.[12] At the last

11. Widerker, as will be recalled, initially claimed that 'the option of continuing to deliberate is *practically* equivalent to deciding not to keep the promise' (2006, 170, emphasis added).

12. Cf. Larvor (2010, 507).

> second, he once again develops a desire to keep his promise, verges on deciding to keep it, but *N*'s absence stymies his 'effort' to do the right thing. In the end, he runs out of time to deliberate, misses the bus to the hospital, and thereby breaks his promise.

With an eye toward forestalling the objection that I am begging the question against Widerker's stipulation that Jones (in BMF) is unable to decide to keep his promise, it is worth making explicit two implicit assumptions on which I proceed in this case: (i) that one can adopt a favourable attitude toward doing *V* without having decided to do *V*, and (ii) that one can have a desire to do *V* without having decided to do *V*—for example, I can have a desire to go on holiday without having decided to go on holiday. Although Jones, then, in BMF*, adopts a favourable attitude toward keeping his promise and forms a desire to keep it, he does not decide to keep it.

Now let us address the question whether Jones is blameworthy in the (counterfactual) scenario described in BMF*. Intuitions, while not inviolable, are an indispensable starting point, and I submit that, intuitively, he is not blameworthy. Had he not been frustrated by circumstances beyond his control—i.e., by the absence of *N*—he would have decided to keep his promise. Indeed, but for the absence of *N*, he would have stopped deliberating when he formed the desire to keep it and, at that point, decided to keep it. To be sure, moral responsibility is not an all-or-nothing affair: the moral disvalue (or value) of an action, and hence the degree of blameworthiness (or praise) we attribute to the agent who performed the action, is usually a matter of degree.[13] Since (we may suppose) he momentarily contemplated breaking his promise in the first place, therefore, Jones is, arguably, *somewhat* blameworthy in the counterfactual sequence in which he continues deliberating. But if he is merely somewhat blameworthy, then his continuing to deliberate is a morally significant alternative; and this is because had he simply decided outright to break his promise, he would have been *wholly* blameworthy. Continuing to deliberate, then, mitigates his blameworthiness (even if it does not exempt him from it altogether).

It bears underscoring, however, that we cannot rule out the possibility that Jones never so much as entertains the idea of breaking his promise. More fully, we cannot rule out the possibility that his entire deliberative process (prior to *t*) consists of his repeatedly verging on deciding outright to keep it (only to be foiled, time and again, by the absence of *N*). In other words, Jones has an alternative that exempts him from blame altogether.

13. I discuss this further in section 3 below.

Friends of PAP, in any event, will insist that we cannot rule out this possibility, at least not without begging the question against PAP. Either way—whether he is somewhat blameworthy or not blameworthy at all—Jones's continuing to deliberate, in the way described in BMF*, is *not* morally equivalent to his deciding to break his promise.

We can augment the case for this conclusion by formulating a converse scenario in which Jones is unable to decide to *break* his promise. So suppose that, shortly before beginning to deliberate, he undergoes a neurological change as a result of which one of the neurological causally necessary conditions for his deciding to break his promise does not obtain. Further suppose that, while deliberating, he adopts a favourable attitude toward breaking it and, as a result, forms a desire to break it. Finally, suppose that a back-and-forth similar to the one described in BMF* transpires, such that, in the end, Jones runs out of time to deliberate and thereby keeps his promise. Here Jones is not praiseworthy because he continued deliberating despite wanting to break his promise. But this situation parallels the situation described in BMF*, just in converse, so if Jones is not praiseworthy in this scenario, then, by parity of reasoning, he is not blameworthy in BMF*.[14]

Note that it does not matter—and with this Widerker would agree—that Jones's acts as described in BMF* occur covertly within consciousness. Mental acts, such as deciding, concentrating, pondering, undertaking, imagining, forming an intention, mulling over, etc., are still acts, things we do. Moreover, for libertarians, mental acts constitute the *loci* of moral responsibility, so any non-question begging argument against PAP cannot appeal to the distinction between mental acts and non-mental acts.

The following principle bolsters the intuition that Jones is not blameworthy in BMF* (where '*V*' is an action that *S* is obligated to perform at *t*):

> (NB) If, through no fault of her own, *S* is prevented from doing *V* despite doing everything in her power to do *V*,[15] then S is not blameworthy for failing to do *V*.[16]

14. There are asymmetries between moral praise and moral blame (for discussion, see Widerker 1991), but, as far as I can tell, they do not affect the point made here.

15. Does it matter that, in BMF*, Jones is not *aware* that he is doing everything in his power to do *V*? I don't see that it does. In many circumstances, to be sure, being absolved of blame (or being held blameworthy, or whatever) will no doubt depend on one's being aware of one's actions, but in the circumstances under consideration it does not.

16. The 'through no fault of her own' clause is important. If *S* is antecedently responsible for bringing about whatever prevented her from being able to do *V*, then arguably she is blameworthy for failing to do *V* even

This principle is reasonable. Indeed, it is not clear how one might argue for it, beyond simply adducing examples involving agents who, by doing everything in their power to perform obligatory actions, avoid blame for failing (through no fault of their own) to perform them. But NB is itself undergirded by the celebrated Kantian maxim that 'ought' implies 'can'. Why does an agent, by doing everything in her power to perform an obligatory action, avoid blame for failing (through no fault of her own) to perform it? A natural answer is that she avoids blame because she *cannot* perform the action in question: circumstances beyond her control have rendered her powerless to do what she is obligated to do. Thus, the principle that 'ought' implies 'can' supports NB.[17]

At the risk of being slightly tedious, I will state the main argument of this section in another way. On Widerker's view, Jones is unable to avoid deciding to break his promise. Conditions are such that either he will decide to break it on his own or he will run out of time to deliberate (and thereby do something morally equivalent to deciding to break it). There is no third possibility, at least none that Jones can actualize by his own efforts and that will exempt him from blame. I have advanced considerations suggesting that this is false, that there is a third possibility, one that Jones can actualize by his own efforts and that will exempt him from blame: he can continue deliberating in such a way that he adopts a favourable attitude toward keeping his promise and, as a result, forms a desire to keep it. If this desire persists, or keeps recurring, such that he continues deliberating until the moment he runs out of time to deliberate, then he has avoided blame, since he has done everything in his power to decide to keep his promise (only to be thwarted, unbeknownst to him, by a neurological deficit). We have, then, a morally significant alternative: Jones continues deliberating in such a way that he avoids blame.

Of course, the 'in such a way' clause here is important. If Jones continues deliberating for purely selfish reasons—if, for example, he continues deliberating solely because he is worried that his uncle will not bequeath to him his valuable coin collection if he breaks his promise—then his continuing to deliberate does not exempt him from responsibility.[18]

---

if she did everything in her power to do *V*.

17. Kant's maxim has been contested, of course, but it is intuitively appealing, theoretically and explanatorily important—e.g., it explains why it makes no sense to say of the wheelchair bound person that he ought to wade into the pond to save the drowning child—and sufficiently widely accepted to justify assuming it here without further argument.

18. Would deliberating in this way exempt him from responsibility for breaking his promise (even if it does not exempt him from responsibility for, well, deliberating in this way, which is itself objectionable)? Perhaps, but not obviously. Suppose that his uncle chastises him for breaking his promise. Intuitively, it will not do

Jones avoids blame only if he continues deliberating in the *appropriate* way, i.e., in a way that exhibits a proper respect for morality (like the way described in BMF*).

To be sure, in the scenario, BMF, that Widerker initially invites us to imagine, Jones simply decides, on his own, to break his promise without forming a desire to keep it. But that is irrelevant, since Jones, if he is libertarianly free, *has it within his power* to perform (mental) actions that exempt him from blame. He can adopt a favourable attitude toward keeping his promise and, as a result, form a desire to keep it. It's just that, were he to form such a desire, he would be unable, due of the absence of *N*, to act on this desire and decide to keep it. Thus, focusing on what Jones does in the non-actual world in which he forms a desire to keep his promise is perfectly legitimate, since the goal is to illustrate what Jones has the *option* of doing in the actual world in which he does not form this desire; and in the actual world he does have the option of continuing to deliberate in the appropriate way (and hence of forming this desire).

It would seem, then, that the alternative of continuing to deliberate cannot be ruled out by adding a time constraint: even under such a constraint, Jones has the option of deliberating in such a way that he avoids blame. Consequently, Widerker has failed, so far, to disprove PAP: a friend of the principle is entitled to ground (in part at least) Jones's responsibility for breaking his promise on this alternative. I turn now to the issue of whether the option of continuing to deliberate can be ruled out by stipulation.

2.2 The option of continuing to deliberate cannot be ruled out by stipulation

Anticipating no uncertainties or objections, Widerker simply asserts, without argument or elaboration, that the alternative of continuing to deliberate can be ruled out by stipulation. But a strong case can be made that this alternative cannot be ruled out this way: on any (reasonable) understanding of it, the stipulation either disregards Jones's libertarian freedom, which libertarian supporters of PAP will assume he possesses, undercuts his moral fitness, or rests on otherwise dubious assumptions.

Let us first construe the stipulation as the claim that Jones is an obstinate, unreasonable person, a person who, like Frankfurt's (1969, 831) Jones_1, 'does what he has once decided to do no matter what happens next and no matter what the cost' (cf.

for Jones to reply by saying, 'Well, I broke my promise, but I'm not blameworthy for breaking it because, in the process of deliberating about whether to keep it, it occurred to me that, if I break it, you might not bequeath to me your coin collection; and this caused me to verge on deciding to keep it.' Jones's reply runs aground, and the reason, it would seem, is that his deliberative process (regarding whether he should keep his promise) did not show the proper respect for morality.

McKenna 2003, 211). In other words, let us construe it as the stipulation that Jones is simply not the sort of agent who, in the circumstances, would continuing deliberating. He knows his time is short and, given his abnormally headstrong personality, he would not let this opportunity to spend time with Mary pass him by.

This will not do, since even if Jones is unreasonable, he has the *option*, if he is libertarianly free, of continuing to deliberate (in the appropriate way). Unreasonableness about moral matters does not entail *ignorance* of moral matters. Jones, despite his assumed obstinacy, believes that he should (and can) discharge his obligation to keep his promise.

Moreover, and relatedly, it is conceptually objectionable, even assuming that his unreasonableness is psychopathological, that Jones lacks the ability to continue deliberating. Psychopathology, contrary to its conception in philosophy, does not deprive its subjects of the ability to do otherwise. As Pickard (2015), through a painstaking analysis of disorders of agency, has recently shown,[19] 'psychopathology does not strip people of free will….[A]ddicts, agoraphobics, kleptomaniacs, neurotics, obsessives, psychopathic serial murderers, and, further, patients diagnosed with disorders whose symptoms include impulsive behaviour, such as personality disorders, eating disorders, and paraphilias, have the ability to do otherwise: it is possible for them to refrain from performing the actions constitutive of the disorder' (Pickard 2015, 137). One would have thought this was obvious; after all, people with addictions and disorders patently do refrain from performing the actions constitutive of their addiction or disorder every day, in their thousands. But a sizeable number of philosophers[20] seem to think that if an addiction or disorder has a sufficiently powerful grip on a person, the person is incapable of doing otherwise. This is simply false, however. Although the prognosis for addiction (for example) is, in some cases, bleak, even daily, long-term users of extremely addictive drugs like heroin, methamphetamines, and nicotine often quit.[21] To be sure, people do sometimes lack the capacity for behavioural control, but in these cases their behaviour is 'due to the effect on executive function of their emotional or physical state, in which case [it] approximates an automatic reflex rather than being…an action' (*ibid*, 156). The claim being made is that there are no real-world examples of psychologically determined

19. Cf. Hyman (2007), Glannon (2008), and Heyman (2009).

20. Pickard (2015, 136) provides several examples.

21. Here is Heyman on some of the addicts featured in his book on addiction: 'Scott was a daily methamphetamine user, then a daily heroin user; Jessie was doing cocaine at work and at home, and Patty used cocaine for fifteen years. Yet they quit' (2009, 66).

action: we have no examples of genuine actions over which agents lack control.[22] Or so the libertarian friend of PAP may plausibly argue.

But, too, if the opponent of PAP wishes to maintain that Jones is both unreasonable *and* ignorant of moral matters, or that his unreasonableness about moral matters is a consequence of his (unwilful, blameless) ignorance of moral matters, then one can reasonably rejoin that he is morally incompetent and hence not blameworthy for the wrong choices he makes by virtue this ignorance. Inculpable ignorance is exculpatory, which is why we do not (or at least should not) punish mentally disabled persons who, unbeknownst to them, act wrongly.

At any rate, Widerker himself would, I suspect, resist a reply on his behalf that appealed to Jones's being morally ignorant. Moya, in his defence of PAP against Widerker's attack on the principle, argues that Jones is not blameworthy because, due to the absence of *N*, his capacity for practical reasoning is impaired, and therefore, 'he cannot decide to keep his promise, no matter how strong or decisive the moral reasons for this decision may appear to him' (Moya 2007, 483). Widerker, in his rejoinder to Moya, maintains that even if Jones's capacity for practical decision-making is impaired, as it patently is in BMF, it does not follow that his capacity for practical reasoning is impaired. 'Moya's mistake,' he argues,

> stems from thinking that an agent's sound capacity for practical reasoning entails a capacity for decision-making on the basis of reasons. An agent's capacity for practical reasoning requires that he be reasons-responsive in the sense of being able to respond differentially to reasons. However, this ability need not be cashed out in terms of *decision-making*, i.e., in terms of the different decisions the agent would be able to make (on the basis of reasons) in different circumstances. It can be cashed out in terms of his being able to have/form reasonable *beliefs* as to what he would do in various circumstances, including *beliefs* about which decisions he would make when being presented with different reasons for acting…[and] this is an ability that in [BMF] Jones *had*. (2009, 92–3)

I concur with this, and I invite the reader's concurrence with it. But note that if, on the one hand, we assume that Jones is reasons-responsive in the way described here, then we cannot, on the other hand, stipulate that he is morally ignorant, since being reasons-

22. Cf. Alvarez (2009).

responsive in this way requires being knowledgeable, at least roughly, about which reasons are pertinent (morally speaking) in the circumstances, which reasons are to be rejected, and so on.

Let us now construe the stipulation that Jones lacks the alternative of continuing to deliberate as the claim that BMF can be revised so that *N* is a causally necessary condition, not only for Jones to decide to keep his promise, but also for him to adopt a favourable attitude toward keeping it (and hence for him to form a desire to keep it). In short, let us construe it as the stipulation that *N*'s absence renders Jones incapable of deliberating further.

This will not do, either, since such a stipulation—quite apart from the fact that it concedes that BMF is not a counterexample to PAP—undermines Widerker's response to Moya's objection that, because his capacity for practical reasoning is impaired, Jones is not blameworthy. As we have seen, Widerker, in order to deflect Moya's objection, argues that, despite the fact that Jones's capacity for practical decision-making is impaired, his capacity for practical reasoning is not, since he is able to respond differentially to reasons and is thereby able to form reasonable beliefs as to what to do in various circumstances. However, if Jones cannot so much as adopt a favourable attitude toward keeping his promise, then he is not suitably reasons-responsive, since to adopt a favourable attitude toward *V* involves forming the belief that *V* is to be favoured. But Jones, by our assumption, is unable to form such a belief.

Elaborating, suppose that Jones's neurological deficit has rendered him incapable of even forming the belief that he should keep his promise. Further suppose that, while deliberating about whether to keep it, he reflects on the various (moral) reasons for keeping it (e.g., that breaking it will have bad consequences, that virtuous people keep their promises, etc.). It seems plainly false that although he is incapable of forming the belief (on the basis of these reasons) that he should keep his promise, he possesses the ability to form reasonable beliefs as to what to do in various circumstances. Trivially, the belief that he should keep his promise is reasonable, and, by our assumption, he is unable to form this belief. By Widerker's own criteria, then, of what constitutes a sound capacity for practical reasoning, Jones lacks such a capacity. But then it follows that the circumstances of the revised IRR-situation (in which *N*'s absence renders Jones incapable of even forming the belief that he should keep his promise) are such that Jones is not fit to be held responsible, since he fails to meet a requirement for responsibility of the reasons-responsiveness account of responsibility to which Widerker subscribes. Speaking more generally, to strip an agent of his ability to form reasonable beliefs as to what to do in various circumstances is effectively to strip him of his capacity to act *rationally*. But

responsibility requires this capacity, which is why a person who acts wrongly because this capacity is impaired, say because of dementia, is not blameworthy.

Finally, let us interpret the stipulation that Jones lacks the alternative of continuing to deliberate as the claim that a condition on his successfully keeping the promise is that, at the time he makes it, he must firmly commit to keeping it and agree not to subsequently deliberate about whether to keep it, say because his uncle has made this a term of their 'agreement'.[23] Call this condition *C*. If Jones agrees to *C*, then after he makes the promise, he cannot deliberate about whether to keep it without doing something morally equivalent to deciding to break it.

This interpretation of the stipulation, however, rests on an assumption that has no demand on our acceptance, namely, that it is possible to do something morally equivalent to deciding to break a promise simply by deliberating about whether to keep it. Suppose that Jones agrees to *C*, makes the promise to his uncle, and subsequently contemplates breaking it. But further suppose, counterfactually, that he has not suffered a neurological malfunction and that, in the end, he visits his uncle in the hospital as promised. Even if (contrary, as I see it, to fact) his uncle would be justified in reproaching him for contemplating breaking the promise, he would not be justified in accusing him of doing something morally equivalent to deciding to break it, since he did no such thing! Although our actions are under our voluntarily control, our motives are not: we can foster virtuous motives, but we cannot directly produce them in ourselves from one moment to the next.[24] And nor can we always control our thoughts, which sometimes occur without discernible impetus. One does not do something morally equivalent to deciding to break a promise if one wavers momentarily in one's commitment to it, say because one's non-virtuous motives or spontaneous thoughts have led one to contemplate breaking it.

3. An objection rebutted

In this section I confront an important objection that might be levelled against my argument that BMF does not constitute a counterexample to PAP.

3.1 The Objection

Reflection on the epistemic dimension of Jones's situation in BMF reveals that continuing to deliberate in the appropriate way is insufficiently *robust*, to use Fischer's

23. I thank an anonymous reviewer for suggesting this interpretation of the stipulation.

24. See, e.g., Ross (1930, 5).

(e.g., 1999) well-known terminology, to qualify as a morally significant alternative; it is too 'flimsy and exiguous' to ground an attribution of moral responsibility. Jones is cognitively insensitive to the fact that continuing to deliberate in the appropriate way exempts him from blame. Widerker has adduced an example, call it POTION, in which near Jones is a cup of water that, unbeknownst to him, contains a sleeping potion. It is up to Jones whether he drinks the water, but this alternative is irrelevant to his moral responsibility, since he is cognitively insensitive to the fact that by drinking the water he can avoid breaking his promise.[25] 'In such a situation,' Widerker avers, 'it would be counterintuitive to ground Jones's culpability for D(– K) in the fact that he did not avail himself of the alternative possibility of drinking the said cup of water' (2006, 175).[26] Similarly, says our objector, it would be counterintuitive to ground Jones's culpability for D(– K) in the fact that he did not avail himself of the option of continuing to deliberate. Succinctly put, the option of continuing to deliberate fails to meet the requirements for a morally significant alternative. As I have noted, Widerker defines a morally significant alternative as follows:

> (MSA_1): [A]n act *V** [is] a morally significant alternative to a morally wrong act *V* that an agent *S* performs at *t* if and only if: (i) *V* and *V** are incompatible, (ii) *S* is aware (or should be aware) that if he performs *V** at *t*, he will not be acting in a morally wrong way, (iii) *S* truly believes that it is within his power to perform *V** at *t*, and (iv) in the circumstances, *V**-ing at *t* would be regarded as a reasonable way for *S* to avoid *V*-ing at *t*. (2006, 175) [27]

Observe that the option of continuing to deliberate fails to meet the second (and possibly the fourth) of these requirements.

Reply: We may dispute the assumption that Widerker's definition of a morally significant alternative (MSA_1) is adequate. Our objector assumes that it is, and then

25. See also Pereboom (2003, esp. 187–8).

26. Strictly speaking, drinking the water is a morally significant alternative, since breaking the promise has moral value, and if *V* has moral value, then any action with a different moral value will be morally better or morally worse than *V* and thereby be significant. As McKenna has pointed out, however, 'within the very wide spectrum of courses of action with differing moral weights or values, only some in a deliberative context will be relevant to *competent moral deliberation and agency*' (2003, 207). Drinking the water is not, in this latter sense, a morally significant alternative.

27. I have altered Widerker's notation slightly.

claims, in essence, that any alternative that fails to satisfy its requirements has to be jettisoned. But why should we accept this line of reasoning? Why should we not assume, instead, that NB is true and that continuing to deliberate is actionally accessible to Jones, and then claim that since MSA_1 excludes this alternative, MSA_1 has to be revised? After all, MSA_1 is based solely on its purported intuitive resonance,[28] and since the alternative of continuing to deliberate seems intuitively to be morally significant, it is reasonable to conclude that MSA_1 requires revision.

These reflections suggest that we need to construct a definition of a morally significant alternative that, while it excludes from its extension alternatives like the one described in POTION, includes in its extension alternatives like the one described in BMF*. This can be accomplished by augmenting MSA_1 as follows:

> If, through no fault of her own, *S* is prevented from doing *V** despite doing everything in her power to do *V**, then *S*'s doing everything in her power to do *V** is a morally significant alternative to doing *V*.

As the reader will notice, this is a variation on NB. If this clause is appended to MSA_1, continuing to deliberate in the appropriate way is a morally significant alternative (where continuing to deliberate constitutes Jones's doing everything in his power to decide to keep his promise), whereas alternatives like the one described in POTION are not.

But there is another reason to reject MSA_1 that is independent of the fact that it conflicts with my contention that the option of continuing to deliberate is a morally significant alternative. As previously discussed, not all wrong actions are equally bad (just as not all right actions are equally good), so the degree of blameworthiness we attribute to an agent who has performed a wrong action will depend on the wrongfulness of the action. While Jones, then, is perhaps blameworthy for momentarily contemplating breaking his promise in the first place,[29] he is not *as* blameworthy as he would have been had he decided outright to break it. Thus, he could have acted appreciably less badly than he did (while still doing something blameworthy), something he knew (we may assume). But MSA_1 cannot capture this intuitive fact, for if MSA_1 is correct, then an

28. It is also noteworthy that there are competing definitions in the literature. McKenna (2003, 209) and Pereboom (2003, 188), for example, have offered definitions that Widerker finds unsatisfactory.

29. Though, again, we cannot rule out the possibility that, while he was deliberating, Jones never so much as entertains the idea of breaking his promise, that his whole deliberative process (prior to *t*) consists of his repeatedly verging on deciding outright to keep it (only to be foiled by the absence of *N*). Since we cannot rule out this possibility, Jones has an alternative that allows him to avoid blame altogether.

agent has to be aware (or should be aware) that if he performs *V** (a morally significant alternative to a wrong act *V*), 'he will not be acting in a morally wrong way'. With this in mind, consider the following slightly modified version of MSA_1:

> (MSA_2): [A]n act *V** [is] a morally significant alternative to a morally wrong act *V* that an agent *S* performs at *t* if and only if: (i) *V* and *V** are incompatible, (ii) *S* is aware (or should be aware) that if he performs *V** at *t*, he will be acting less badly than if he does *V*, (iii) *S* truly believes that it is within his power to perform *V** at *t*, and (iv) in the circumstances, *V**-ing at *t* would be regarded as a reasonable way for *S* to avoid *V*-ing at *t*.

This version of a morally significant alternative retains what is attractive about MSA_1, and it includes in its extension the alternative of continuing to deliberate in the appropriate way (while excluding alternatives like the one described in POTION), but it is superior to MSA_1 because, unlike MSA_1, it accommodates the intuition that a morally significant alternative to a wrong action need not be an action that the agent believes, or should believe, is itself not morally wrong.

In any case, the strategies I have adumbrated—viz., augmenting MSA_1 or replacing it with MSA_2—are reasonable, which means that the PAP enthusiast need not acquiesce to the claim that since the alternative of continuing to deliberate fails to meet the requirements of a morally significant alternative as specified in MSA_1, the alternative of continuing to deliberate is not morally significant.

4. Conclusion

I conclude that BMF is not an IRR-situation and, consequently, that Widerker has failed to generate a convincing argument against PAP. Attempts to overturn PAP by dint of counterexample have not fared well—certainly there is nothing resembling a consensus regarding whether the myriad fanciful counterexamples on offer succeed—and, as it turns out, Widerker's proposed counterexample, as ingenious and initially compelling as it is, is just another putative counterexample in a long line of putative counterexamples that, on closer inspection, fails to demonstrate that the principle is false.

References

Alvarez, M. 2009. "Actions, Thought-Experiments and the 'Principle of Alternate Possibilities.'" *Australasian Journal of Philosophy* 87: 61–81.

Fischer, J. M. 1999. "Recent Work on Moral Responsibility." *Ethics* 110: 93–139.

Fischer, J. M. 2010. "The Frankfurt Cases: The Moral of the Story." *Philosophical Review* 119: 315–36.

Frankfurt, H. 1969. "Alternate Possibilities and Moral Responsibility." *Journal of Philosophy* 66: 829–39.

Ginet, C. 1990. *On Action*. New York: Cambridge University Press.

Ginet, C. 1996. "In Defense of the Principle of Alternative Possibilities: Why I Don't Find Frankfurt's Argument Convincing." *Philosophical Perspectives* 10: 403–17.

Haji, I. and McKenna, M. 2004. "Dialectical Delicacies in the Debate about Freedom and Alternative Possibilities." *Journal of Philosophy* 101: 299–314.

Glannon, W. 2008. "Moral Responsibility and the Psychopath." *Neuroethics* 1: 158–66.

Heyman, G. M. 2009. *Addiction: A Disorder of Choice*. Harvard: Harvard University Press.

Hyman, S. E. 2007. "The Neurobiology of Addiction: Implications for Voluntary Control of Behavior." *The American Journal of Bioethics* 7: 8–11.

Kane, R. 1996. *The Significance of Free Will*. New York: Oxford University Press.

Larvor, B. 2010. "Frankfurt Counter-Example Defused." *Analysis* 70: 506–8.

McKenna, M. 2003. "Robustness, Control, and the Demand for Morally Significant Alternatives: Frankfurt Examples with Oodles and Oodles of Alternatives." In *Moral Responsibility and Alternative Possibilities: Essays on the Importance of Alternative Possibilities*, edited by Widerker and McKenna, 201–17. Aldershot, UK: Ashgate.

Mele, A. and Robb, D. 1998. "Rescuing Frankfurt-Style Cases." *Philosophical Review* 107: 97–112.

Moya, C. J. 2007. "Moral Responsibility without Alternative Possibilities?" *Journal of Philosophy* 104: 475–86.

Pereboom, D. 2003. "Source Incompatibilism and Alternative Possibilities." In *Moral Responsibility and Alternative Possibilities: Essays on the Importance of Alternative Possibilities*, edited by Widerker and McKenna, 185–99. Aldershot, UK: Ashgate.

Pickard, H. 2015. "Psychopathology and the Ability to Do Otherwise." *Philosophy and Phenomenological Research* 90: 135–63.

Ross, W. D. 1930. *The Right and the Good*. Oxford: Clarendon Press.

van Inwagen, P. 1983. *An Essay on Free Will*. Oxford: Clarendon Press.

Widerker, D. 1991. "Frankfurt on 'Ought Implies Can' and AlternativePossibilities." *Analysis* 51: 222–4.

Widerker, D. 1995. "Libertarianism and Frankfurt's Attack on the Principle of Alternative Possibilities." *Philosophical Review* 104: 247–61.

Widerker, D. 2000. "Frankfurt's Attack on the Principle of Alternative Possibilities: A Further Look." *Philosophical Perspectives* 14: 181–201.

Widerker, D. 2003. "Blameworthiness and Frankfurt's Argument against the Principle of Alternative Possibilities." In *Moral Responsibility and Alternative Possibilities: Essays on the Importance of Alternative Possibilities*, edited by Widerker and McKenna, 53–73. Aldershot, UK: Ashgate.

Widerker, D. 2005. "Blameworthiness, Non-robust Alternatives, and the Principle of Alternative Expectations." *Midwest Studies in Philosophy* 29: 292–306.

Widerker, D. 2006. "Libertarianism and the Philosophical Significance of Frankfurt Scenarios." *Journal of Philosophy* 103: 169–87.

Widerker, D. 2009. "A Defense of Frankfurt-friendly Libertarianism." *Philosophical Explorations* 12: 87–108.

Widerker, D. and Goetz, S. 2013. "Fischer against the Dilemma Defence: The Defence Prevails." *Analysis* 73: 283–95.

Widerker, D. and McKenna, M. (eds.). 2003. *Moral Responsibility and Alternative Possibilities: Essays on the Importance of Alternative Possibilities*. Aldershot, UK: Ashgate.

Journal of Cognition and Neuroethics

Reverse Inference and Mind-Brain Identity

Yakir Levin
Ben-Gurion University of the Negev

Itzhak Aharon
Interdisciplinary Center (IDC) Herzliya

Biographies

Itzhak Aharon (Gingi) is a senior lecturer at the Interdisciplinary Center (IDC) Herzliya, Israel. His research focuses on the neurobiology of motivation and decision making (neuroeconomics).

Yakir Levin is a senior lecturer in philosophy at Ben-Gurion University of the Negev, Israel. His research interests include early-modern philosophy, analytic metaphysics, and philosophy of mind.

Acknowledgements

Earlier versions of the paper were presented at two conferences: (1) The Aims of Brain Research: Scientific and Philosophical Perspectives, co-organized by the Safra Center for Brain Sciences (ELSC) and the Edelstein Center, Hebrew University, the Cohn Institute, Tel Aviv University, and the Van Leer Jerusalem Institute; (2) The Nathan Stemmer Memorial Colloquium on the Philosophy of the Science of Morality co-organized by the Edelstein Center and the Center for Moral and Political Philosophy, Hebrew University. Thanks are due to the following participants in these events for their helpful comments and suggestions: Jean-Pierre Changeux, Meir Hemmo, Eva Jablonka, Tom Polger, and Adina Roskies (first conference); David Enoch, Edouard Machery, Boaz Miller, and Chandra Sripada (second conference). In developing the main argument of the paper, we deploy some material from Sects. 2 and 3 of Levin and Aharon 2011 in a significantly revised form.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). September, 2015. Volume 3, Issue 2.

Citation

Levin, Yakir, and Itzhak Aharon. 2015. "Reverse Inference and Mind-Brain Identity." *Journal of Cognition and Neuroethics* 3 (2): 23–45.

Reverse Inference and Mind-Brain Identity

Yakir Levin and Itzhak Aharon

Abstract
Reverse inference is a widespread procedure of reasoning from patterns of brain activation to the engagement of specific mental processes. One of the main attractions of reverse inference is the apparent possibility it opens for exceeding the limits of behavior based procedures in psychology. Underlying this motivation is an implicit assumption of mind-brain identity according to which behavior does not play a constititive role in the classification of mental kinds. A widely accepted consideration, however, against mind-brain identity is that while identity is a one-one relation, there appears to be mounting evidence that the mind-brain relation is one-many. In this paper we examine three recent strategies for responding to this consideration, positing that two of them fail, while the third may be successful but at the cost of giving the behavioral criteria of mental kinds a constitutive role to play in the classification of brain kinds. For this reason, the third strategy does not yield an account of mind-brain identity capable of grounding a positive answer to the question of whether reverse inference can exceed the limits of behavior based procedures in psychology. It follows that philosophical defenses of mind-brain identity may be useless for the purposes of science. Nevertheless, they are not irrelevant to science, since their failure to ground a specific scientific research strategy such as reverse inference may constitute an important negative lesson concerning this strategy.

Keywords
Reverse Inference, Mind-brain Identity, Multiple Realization, Degeneracy, Pleiotropy

1. At the Intersection of Cognitive Neuroscience and Philosophy

1.1 Forward vs. Reverse Inference

The classic strategy employed by neuroimaging researchers has been to manipulate a specific psychological function and identify the localized effects of this manipulation on brain activity. This has been referred to as "forward inference" (Henson 2005), and is the basis for a large body of knowledge that has been derived from neuroimaging research. However, since the early days of neuroimaging, there has also been a desire to reason backward from patterns of activation to the engagement of specific mental processes. This has been called "reverse inference" (Poldrack 2006), and often forms much of the reasoning observed in the discussion section of neuroimaging papers (under the guise of "interpreting the results") (Poldrack 2011, 692).

Having been so widespread in the neuroimaging literature – an "epidemic" as one writer has described it (Poldrack 2006) - reverse inference's common use has not gone unchallenged (Aguirre 2003; Poldrack and Wagner 2004; Henson 2005; Page 2006; Poldrack 2006; Christoff and Owen 2006; Poldrack 2008; Harrison 2008; Van Horn and Poldrack 2009; Bourgeois-Gironde 2010; Poldrack 2011; Fox and Friston 2012). This has led to its becoming a bad name in some quarters (Poldrack 2011, 692). However, short of undermining it, these criticisms have paved the way for a more cautious and sophisticated use which, e.g., focuses on patterns of activation rather than localized blobs, utilizes broader and more comprehensive fMRI databases, makes use of high-performance computer clusters, deploys improved techniques of statistical analysis etc. (Poldrack 2011 and 2012; Hutzler 2014). These criticisms, moreover, have certainly not deterred neuroimaging researchers – especially in areas such as neuroeconomics and social neuroscience in which the underlying mental processes are less well understood - from regarding reverse inference as a fundamentally important research tool (see, e.g., Young and Saxe 2009) – "the *sine qua non* of inference in neuroeconomics" as one researcher has put it (Harrison 2008, 535).

1.2 Behavior Exceeding Reverse Inferences

One reason that reverse inference has been considered so important in neuroeconomics is that it is often not possible to determine the correctness of cognitive theories adduced in this field solely on behavioral basis. Thus, consider the well-known tendency of consumers to behave "impatiently" today but to prefer/plan to act "patiently" in the future. For example, someone offered the choice between receiving $10 today and $11 tomorrow is likely to choose the immediate option. However, if asked today to choose between $10 in a year and $11 in a year and a day, the same person is likely to prefer the slightly delayed but larger amount. One hypothesis that has been advanced to explain this phenomenon is that it reflects the operation of two fundamentally different mechanisms, one affective, which heavily values the present and steeply discounts all future opportunities, and the other deliberative, which discounts options more consistently across time. However, it has not been possible to provide evidence for separate mechanisms from behavioral data alone, or to motivate them on the basis of purely theoretical considerations (Sanfey, Loewenstein, McClure, and Cohen 2006, 113).

It has been, therefore, the hope of neuroeconomists that neurobiological data could play here the evidential role that behavioral data does not, perhaps even cannot, play. And indeed, as a recent fMRI study has shown, choices involving the option of

an immediate reward actively engage the ventral striatum, as well as the medial and orbitofrontal areas – areas rich in dopaminergic innervation, and consistently associated with the evaluation of reward (McClure, Laibson, Loewenstein, and Cohen 2004). In addition, this study has shown that both choices involving the option of an immediate reward and those involving the option of a delayed reward consistently involve areas of the frontal and parietal cortex commonly associated with more abstract forms of reasoning and planning. And this has been taken to corroborate the aforementioned hypothesis – viz., that the difference between the short-term and long-term choices at issue reflects the operation of two different cognitive mechanisms, one affective and the other deliberative (Sanfey, Loewenstein, McClure, and Cohen 2006, 113; Camerer, Loewenstein, and Prelec 2005, § 5.1; but cf. Kable and Glimcher 2007; Bernheim 2009, § 1.5.1).

In like manner, reverse inference has been employed for dissociating social decision making of the sort often employed in behavioral economics games (e.g., trust game, ultimatum game, prisoner's dilemma game, etc.), and decision-making concerning non-social stimuli or agents, where this cannot be done on a behavioral basis (Lee and Harris 2013). Thus, it has been shown that attributions of behavioral traits to human agents on the basis of their observed behavioral patterns, rely on a distinct set of brain regions, including the medial prefrontal cortex (MPFC) – a region known to be active when value signals in a social context are created – and the superior temporal sulcus (STS). However, when the agents are anthropomorphized objects, although the same patterns of behavior as those manifested by human agents lead to the attribution of the same behavioral traits, the underlying pattern of brain activity is different. Specifically, attributions for objects do not engage MPFC but rather STS and the bilateral amygdala. And this has been taken to suggest that social and non-social decision making involve different cognitive mechanisms.

Another example of a use made of reverse inference in the absence of clear behavioral indications concerns decision making under uncertainty (Huettel et al. 2006). Thus, preferences for risk (uncertainty with known probabilities) have been shown to correlate with activation of the posterior parietal cortex. In contrast, preferences for ambiguity (uncertainty with unknown probabilities) have been shown to correlate with activation of the lateral prefrontal cortex. And this has been taken to indicate that, contrary to standard assumptions, decision making under ambiguity does not represent a special, more complex case of risky decision making; instead, these two forms of uncertainty involve distinct cognitive mechanisms.

Yet another example of a use made of reverse inference in the absence of clear behavioral indications concerns cooperative behavior (Stallen and Sanfey 2013). Thus, reciprocated cooperation in the prisoner's dilemma game and the trust game have been shown to correlate with activity in the ventral striatum and the ventromedial prefrontal cortex, brain regions consistently found to be activated by both social and monetary rewards. Relatedly, viewing the faces of individuals who had previously cooperated in a prisoner's dilemma game, as compared to faces with whom the player had no history, elicited enhanced neural activity in reward-related areas such as the striatum, nucleus accumbens, and orbitofrontal cortex. And this has been taken to suggest that people are motivated to resist the temptation to selfishly accept but not reciprocate favors, by labeling mutual cooperation as rewarding in and of itself – i.e., independent of whatever monetary gain was obtained by the cooperative action.

1.3 Mind-Brain Identity, Multiple Realization, and the Scope of Reverse Inference

A major reason, then, for the attraction of reverse inference is the hope/assumption that it makes it possible to exceed the limits of behavior based procedures in psychology/economics, and achieve what is impossible by these procedures alone. A similar hope/assumption seems to be lurking behind the claims by prominent neuroeconomists such as Camerer, Loewenstein, and Prelec that the study of the brain and the nervous system is "beginning to allow direct measurement of thoughts and feelings," or that neuroscience is making possible the measurement of "basic psychological forces … without inferring them from behavior," or enables "to observe processes and constructs which are typically considered unobservable," or opens for the first time "the 'black box' [sometimes more wittingly called the 'grey box'] … – the human mind" (Camerer, Loewenstein, and Prelec 2004, 572; Camerer, Loewenstein, and Prelec 2005, 10 and 53; Camerer 2007, c38; Camerer 2008a).

Underlying these hopes/assumptions is an implicit assumption of mind-brain identity which takes the behavioral expressions of mental features to be mere contingent indications of these features that do not play a role in their classification into mental kinds (more on this below): If mental features can be inferred on the basis of neurobiological features (such as patterns of activation) when there are no clear - or perhaps even any - behavioral indications for their presence, this would mean that only neurobiological features are constitutive of these mental features and required for their classification into kinds. It would mean, in Francis Crick's (1994, 3) provocative formulation, that "you, your joys and your sorrows, your memories and your ambitions, your sense of

personal identity and free will, are in fact no more than the behavior of a vast assembly of nerve cells and their associated molecules"; that "you are your synapses," in Joseph LeDoux's (2003, 323-324) equally disturbing words; that "mental processes actually are processes of the brain," in Patricia Churchland's (2008, 409) somewhat more prosaic phrasing. So the question of whether reverse inference can exceed the limits of behavior based procedures boils down to the question of the viability of an account of mind-brain identity which does not give behavior a constitutive role in the classification of the brain kinds with which mental kinds are identified. Thus reformulated, however, this question of the scope of reverse inference becomes related to the prototypical argument contra mind-brain identity (Polger 2011, 9; see also Polger 2009, 458), the *multiple realization objection* according to which: *The mind-brain relation is one-many* – i.e., one and the same mental kind can be realized or subserved by distinct kinds of brain structures. But, *the identity relation is one-one* – i.e., if a mental kind is realized or subserved by distinct brain kinds, it cannot be identical with any one of them. Thus, *mind isn't identical to the brain*.

Although there has been a wide consensus among philosophers ever since it was put forward by Putnam and others in the early 1960s that this objection is devastating, it has not gone unchallenged. The first wave of challenges which goes back to the late 1960s, didn't question the very phenomenon of multiple realization of mental kinds by brain kinds. Instead it sought to show that this phenomenon does not really pose a problem to mind-brain identification, or to a reduction of mind to the brain (Bickle 2010, 248-251). A second wave of challenges which arose about a decade ago, has taken a different tack attacking the very claim that mental kinds are multiply realized by brain kinds (Bickle 2010, 253-255; cf. Polger 2008, 538). A major focus of this wave of challenges has been the question of how the realization relation between realized and realizing kinds is to be analyzed, and the related question of how brain realizers are to be classified into kinds. With this focus, however, this challenge and the debate it has evoked, is particularly relevant to the question of the scope of reverse inference in its reformulation as the question of the viability of an account of mind-brain identity which does not give behavior a constitutive role in the classification of the brain kinds with which mental kinds are identified. Do the classificatory criteria of these brain kinds implied by second-wave-responses involve behavioral features? If they do, insofar as the mind-body identity defended by the second-wave-responses is concerned, reverse inference cannot exceed the limits of behavior based procedures.

1.4 Aim, Structure, and Implications

In this paper we address this question. To this end we shall first clarify what we mean by mind-brain identity-cum-reduction (§ 2). We shall then outline a set of considerations that have been widely taken to support the claim that the mind-brain relation is one-many, or that mind is multiply-realized by the brain (§§ 3-4). Finally, we shall outline the second-wave strategies of responding to these considerations, positing that two of them fail, while the third may be successful but at the cost of giving the behavioral criteria of mental kinds a constitutive role to play in the classification of brain kinds (§§ 5-7). Due to this aspect of the third strategy, it does not yield an account of mind-brain identity capable of grounding a positive answer to the question of whether reverse inference can exceed the limits of behavior based procedures (§ 8).

An important implication of this conclusion is that philosophical defenses of mind-brain identity may be useless for the purposes of science. This is not to say, however, that these defenses are irrelevant to science, since their failure to ground a scientific research strategy such as reverse inference, or some of its intended uses-cum-goals, may be an important negative lesson concerning that strategy or those intended uses-cum-goals. So the paper illustrates an interesting link between cognitive neuroscience and philosophy. The other side of the same link, which the paper also illustrates, is the heavy reliance of the second wave of challenges to multiple realization on neuroscientific practices and findings.

2. Mind-Brain Identification and Reduction

2.1 Empirical Identification

Suppose that being in a mental state of a given kind is one and the same thing as being in a brain state of a specific kind. For example, suppose that being in pain is one and the same thing as being in a state of P-fiber excitation. ('P' here stands for the parieto-insular cortex, for which there is growing evidence of its centrality for pain processing in the brain. We also use the term 'P-fiber excitation' in homage to the ubiquitous yet wrong reference that philosophers make in this context to C-fiber excitation: C-fibers are located well outside the brain, existing as a subset of the sensory neurons that project to the spinal cord [Pucceti 1977; Allen, Grau, and Meagher 2009, 129-130].) If such mind-brain identities obtain, they cannot, of course, be certifiable conceptually, or solely from the meanings of the mental terms (e.g., 'pain') and brain terms (e.g., 'P-fiber excitation') that they involve. The concept of pain and the concept of P-fiber excitation are distinct

and independent concepts, and this explains how it is possible for someone to know a lot about pains but nothing about P-fibers or their excitations. If pains are identical with P-fiber excitations, this is an empirical truth whose corroboration depends on elaborate neurobiological research. In like manner, the concept of the morning star (Phosphorus) and that of the evening star (Hesperus) are distinct concepts, and it could only have been discovered by observation and experience that they signify one and the same thing – viz., Venus. And this is also the case with "theoretical identities" in the sciences such as the identity of water and H_2O, or of heat and molecular motion, or of light and electromagnetic radiation. In being empirical, all these identities differ significantly from an identity such as "2=the smallest prime number," which are known a priori, or by merely investigating the meanings of the expressions that they involve.

2.2 Identity of Kinds and Classificatory Criteria

Prior to the discovery of an empirical identity between two kinds, then, these kinds are considered distinct and independent from each other. This being the case, each kind must be associated with a distinct and independent classificatory criteria, or a set of features that determine the things that belong to the kind – e.g., specific intricate behavior in the case of pain, and specific neurobiological features in the case of P-fiber excitation. After the discovery of the identity, however, one of these sets of features may be considered more fundamental than the other, most naturally, the set characterizing the lower level type, a level that is typically considered as of a higher explanatory value ("wherever the bottom is, that is where the real explanations are to be found" is a widely accepted view, not least so in contemporary neuroscience [Craver 2007, 11-15]). Indeed, the set of classificatory criteria considered less fundamental may no longer be considered constitutive or essential to the kind; it may no longer be considered as playing a role in determining the kind, but rather as mere contingent evidence for the occurrence of an instance of the kind.

If such a change in status of classificatory criteria occurs, however, the mere occurrence of something with features of the fundamental classificatory set may be considered sufficient for the occurrence of an instance of the kind, even if it is not accompanied by an occurrence of the other set of classificatory features. Thus, in the 17th century fluidity was considered essential to water (Locke 1975, 4.6.11), and ice, which lacks this feature, was considered a different substance (ibid., 3.6.13). So the transformation of water into ice (or vice versa) was considered then like the transformation of a wooden table to a pile of bits and pieces of wood when the table is

broken apart: Like the latter transformation the former one wasn't considered a change in a mode or state of a thing, but as a transformation of one thing to a completely different thing. After the discovery of the micro-structure of water – viz., H_2O - and the identification of the kind water with the kind H_2O that ensued, this was no longer the case, and something that had not been considered before as looking and behaving like water – viz., ice – began to be considered that way (as water in a solid phase). Indeed, its "non-watery" behavior became part of what is considered standard "watery" behavior in specific circumstances. Relatedly, once water has been identified with H_2O, things that behave and look like water but are not H_2O, cannot be considered as water anymore (Putnam 1975, 223-235).

2.3 From Kind-Identification to Reduction

An identification of one kind of phenomenon (usually of a higher level) with another kind of phenomenon (usually of a lower level) may involve, then, a change in the status of the classificatory criteria that, prior to the identification, were regarded as determining each kind. As a result of the identification only the classificatory criteria associated with the lower level kind may be viewed as constitutive of the unified kind, or as determining it. Insofar as identifications of kinds of phenomena involve such changes in status, they are tantamount to a reduction of the phenomena of the higher level kind to the phenomena of the lower level kind. The question of how common reductions of this sort are in science, may be controversial (Antony 2007, 154-156; Craver 2007, 108-109). Be that as it may, a reduction of this sort may be rather appealing in the mind-brain case. For underlying our common sense view of the mind are Cartesian intuitions which incline us to consider the behavioral criteria of mental states as mere contingent evidence for utterly different things. Cartesians consider these utterly different things as states of a non-material substance. But from here it is but a short step to consider them as states of a material substance, or the brain. So a mind-brain identification-cum-reduction can be considered a materialized version of Cartesianism (Bennett and Hacker 2003, 72).

2.4 The One-One Assumption

Having clarified what we mean by mind-brain identification-cum-reduction, the question immediately arises of what can be the basis for such an identification-cum-reduction? Well, the first and crucial step in this direction is the discovery of systematic mind-brain correlations. Indeed, all serious arguments for mind-brain identity must come up with reasons that would make it compelling, or at least reasonable, to upgrade

mind-brain correlations to identities (Kim 2005, Chap. 5). The mind side of these correlations, however, consists of mental states as these are behaviorally manifested, or can be identified by the behavioral criteria for their ascription: The systematic correlations between pain and P-fiber excitation, say, that are upgraded to an identity between the two, are in fact correlations between pain as it is behaviorally manifested, or genuine pain-behavior, and P-fiber excitation. Thus, a basic hypothesis underpinning the thesis of mind-brain identity is the One-One assumption according to which mind in its behavioral manifestations and brain are systematically – indeed, one-one – correlated – i.e., the same brain structures or processes accompany or subserve the same mental state as this state manifests itself or can be identified by observable behavior. The multiple realization thesis, to which we now turn, is directed against this assumption.

3. Enter Degeneracy and Pleiotropy

A ubiquitous property of biological systems at all organizational levels is *degeneracy*, or the ability of elements that are structurally different to perform the same function or to yield the same output (Edelman and Gally 2001; Greenspan 2001 and 2003). As a biological hypothesis, degeneracy has been posited to explain a number of studies of biological organisms, from yeast to humans, in which striking structural differences at various suborganism-levels appear to have little or no organism-level effects (see the references in Edelman and Gally 2001). To explain, for example, why mutations that eliminate the function of various genes need not cause overt harm (an explanation given in terms of overlapping networks of genes that, given appropriate conditions for gene expression, can produce the same outcome) (see Greenspan 2001 and the references therein).

Being a prerequisite for natural selection as well as a product of this process, degeneracy goes hand in hand with *pleiotropy*– i.e., degenerate structures tend to be versatile in their functions, and usually can be used differently in different processing contexts (Tononi, Sporns, and Edelman 1999; Edelman and Gally 2001; Greenspan 2001; Noppeney, Friston, and Price 2004). Thus, a given gene may subserve function F when activated within one gene network and function G when activated within another. It is also the case that biological structures can exhibit degeneracy within an individual at a time or over time, across individuals of the same species, or even across individuals of different species.

Appropriated to neural systems, degeneracy may be taken to mean that diverse brain structures can subserve or realize the same mental state as this state manifests itself or

can be identified by observable behavior. Thus defined, evidence for neural degeneracy both within and across species appears to abound (Price and Friston 2002; Greenspan 2003; Noppeney, Friston, and Price 2004; Noppeney, Penny, Price, Flandin, and Friston 2006; Aizawa 2007; Aizawa and Gillet 2009a; Aizawa and Gillett 2009b; Richardson 2009). In particular, and although they should be considered with caution, lesion and imaging studies which frequently show that entirely different anatomical areas of the brain can subserve the same cognitive functions, have been widely taken to provide strong evidence for neural degeneracy (Figdor 2010, 428-431; cf. Polger 2008, 461-469; Polger 2011, 10). (Imaging studies sometimes replicate findings of lesion studies [Dronkers, Redfern, and Knight 2000], and sometimes also combine with the latter to provide new data [Price and Friston 2002].)

It is also the case that there is ample evidence for neural pleiotropy: Different cognitive functions appear to be supported by putting the same neural circuits together in different arrangements. In each of these arrangements, an individual brain region may perform a similar information-processing operation (a single "working"), but will not be dedicated to the high-level task to which the arrangement is dedicated as a whole (Anderson 2010a; Anderson and Penner-Wilger 2013).

Neural pleiotropy does not really challenge the One-One assumption. At most, it evinces that the mind-brain correlation that this assumption posits must be - at least partially - between mental states on the one hand and patterns of neural activation rather than local neural blobs on the other hand. Nevertheless, neural pleiotropy presents a serious - although perhaps not insurmountable - challenge to reverse inference (Anderson 2010b, 295; Ramsey et al. 2010; Poldrack 2012, 1217-1218). By contrast, neural degeneracy presents a serious challenge to the One-One assumption. In consequence, it also presents a serious and perhaps insurmountable challenge to behavior exceeding reverse inferences. It is to this dual challenge that we now turn.

4. Degeneracy and Multiple Realization

A particularly natural way of viewing neural degeneracy is as showing that mind is multiply realized by the brain – i.e., that contrary to the One-One assumption, there is a one-many mind-brain correlation (cf. Fidgor 2010). Thus, semantic processing tasks (e.g., picture naming) are presumably subserved by different brain structures, in normal subjects and lesion patients, respectively (Price and Friston 2002). And how can this be taken but as showing that, contrary to the One-One assumption, the same semantic

processing task can be correlated with brain states of different kinds? Well, here are three second-wave responses to this challenge.

5. Going Deeper Down Science's Ontological Hierarchy

Going back to a suggestion made by Paul Churchland (1982) and elaborated more recently by John Bickle (2003) this response concedes that if we leave our neuroscientific understanding at the systems level, then psychoneural multiple realization will appear obvious and unavoidable, especially across species. However, as we move further down levels, into cellular physiology and into the intracellular signaling pathways, commonalities - even across widely divergent species – may be the rule – i.e., molecular pathways that underlie specific cognitive and conscious functions may be the same ones from invertebrates to mammals. Bickle's key psychological example throughout his writings on the topic has been memory consolidation, or the conversion of labile, easily disrupted short-term memories into more durable, stable long-term form. In his view, "the discovery of the shared molecular mechanisms for memory consolidation is probably not some isolated, lucky case, but rather follows from core principles of molecular evolution. As 'molecular and cellular cognition' proceeds, we should expect to discover more evolutionarily conserved examples of unitary molecular 'reducers' of shared psychological kinds. Molecular evolution suggests that they should be the rule, not the exception" (Bickle, 2010, 258).

Somewhat ironically, however, Bickle's key example of his proposal appears to refute this very proposal: As Aizawa (2007) has convincingly shown, the biochemical mechanisms of memory consolidation uncovered by molecular neuroscience reveal substantive multiple realization both across and intra species. And if so, Bickle's defense of the One-One assumption fails.

On top of that, in focusing on a low, cellular level of the brain, Bickle's strategy may be irrelevant to the purposes of neuroimaging strategies such as reverse inference which focus on a high, system level of the brain.

6. Eliminate and Split

Another response to neural degeneracy-cum-multiple realization is to eliminate the multiply realized mental kind by splitting it into uniquely realized kinds each corresponding to one of the different realizing brain kinds (Aizawa and Gillett 2011). A notable example of an actual employment of this strategy is the splitting of memory into distinct sub-types in response to neurobiological dissociation experiments: Once it was

discovered that certain sorts of brain lesions lead to the selective loss of certain memory functions, while certain other sorts of brain lesions lead to selective loss of certain other memory functions, the common assumption that there is a single over-arching type of memory has been replaced by the assumption that there are distinct subtypes of memory, declarative and nondeclarative (Squire 2004). However, as Aizawa and Gillett (2011) show, the elimination-by-splitting strategy cannot be considered a general strategy that is applicable across the board, since that would fail to reflect the nuances of actual scientific practice. Thus, as the science of color vision illustrates, differences amongst realizers may lead only to scientists positing individual differences in the same higher level property rather than to subtyping this property. More strikingly still, and again as the science of color vision illustrates, discovered variations in realizers may lead to no variation in the higher level realized properties. And in that case, not only would the subtyping of color vision by way of its lower level property instances be cumbersome, but using the lower level realizer properties to classify higher level properties into kinds may leave us without higher level theories that can track important regularities or generalizations at the higher level.

7. Reject a Presupposition of the Considerations Pro Multiple Realization

As we saw in § 3, multiple realization is supported by cases of mental functions (e.g., semantic processing tasks such as picture naming) that are subserved by different brain structures. How, rhetorically ask proponents of multiple realization, can this be taken but as showing that the One-One assumption is false?

Yet, an implicit and rather natural assumption that underlies this way of viewing multiple realization – an assumption shared by the strategies of going deeper down science's ontological hierarchy and of eliminate and split - is that diversity in realizer *structure* is tantamount to diversity in realizer *kind*: Unless the diversity in the brain *structures* that subserve picture-naming, means diversity in the *kind* of these brain structures *as* realizers of picture-naming, neural degeneracy in this case would not imply that picture-naming can be subserved by brain states of diverse *kinds*.

Thus formulated, this *structure-determines-kind* assumption, leaves open the question of how different, and in what respects, brain structures have to be in order to belong to different brain kinds (Figdor 2010, Sec. 3). Be the answer to this question as it may (Edelman and Gally 2001; Sullivan 2008, Sec. 2; Aizawa and Gillett 2009b; Figdor 2010, 435), for our purposes suffice it to point out that one can keep to the One-One assumption despite neural degeneracy by way of contesting the structure-determines-

kind assumption. It is to two important responses to the multiple realization objection along this line that we now turn.

7.1 Bechtel's and Mundale's Proposal

The first way of contesting the structure-determines-kind assumption is implied by Bechtel's and Mundale's (1999) seminal attack on the hypothesis that psychological functions are multiply realized. As part of their attempt to show that neuroscientific practice contradicts this hypothesis, they claim that brain mapping practices show that brain taxonomy makes essential use of psychological function. This claim may be contested by way of the very examples that Bechtel and Mundale bring in its support (Aizawa 2009, Section 2). Thus, it seems that brain mapping techniques that involve staining brain tissue in order to highlight different features of brain cells, and discerning differences in structure over the volume of the brain, do not make use of psychological function. For the sake of argument, however, suppose that the classification of brain structures must indeed proceed by appeal to psychological function. In that case, however diverse the brain structures that are correlated with a given mental function are, they can still be considered as forming a single unified kind by the very fact that they are all correlated with the same kind of mental function. Thus, however different the distinct brain structures that were found to subserve picture naming are, by Bechtel's and Mundale's lights these brain structures belong to the same kind due to their correlation with picture naming. Thus, Bechtel's and Mundale's strategy makes it possible to maintain the One-One assumption despite neural degeneracy by way of rejecting the structure-determines-kind assumption. This rescuing move comes with a price, however.

In taking psychological functions to play an essential role in the classification of brain structures with which they are correlated, Bechtel's and Mundale's strategy also gives the behavioral criteria of psychological functions a role in the classification of brain structures. This being the case, this strategy preserves one building block of the mind-brain identity thesis – viz., the One-One assumption – but involves a rejection of another building block of this thesis – viz., the *reductionist assumption* that the behavioral criteria of mental kinds do not play a constitutive role in determining these kinds. If the psychological kind of picture naming is identified with a kind of brain structure, and, as by Bechtel's and Mundale's strategy, this brain kind is determined by the behavioral criteria for picture naming, then these behavioral criteria also play a role in determining the mental kind of picture naming.

7.2 Shapiro's Proposal

The second way of contesting the structure-determines-kind assumption is by way of Shapiro's account of the realization relation (Shapiro 2000, 2004, and 2008). On this account, realizers should be classified on the basis of the causal mechanisms by which they yield the functional types that they realize. Thus, a waiter's corkscrew and a double lever corkscrew are different types of realizers of the function of removing corks from bottles, since they each achieve this function in different ways; each employs a different mechanism in the production of cork removal. In contrast, although steel and aluminum waiter's corkscrews differ in constitution, they should be considered of the same type, since they share the same mechanism for corkscrewing bottles. Similarly, eye types such as the octopus eye and the mammalian eye that focus light onto photoreceptive cells in the same way are considered of the same kind ("camera eye"), even if they achieve these optical characteristics by, say, different molecular structures (Shapiro 2000, 646). In other words, "it is optics that provides the level of description at which a clump of molecules constitutes an eye, and hence it is the science of optics that determines whether two eyes are instances of a single kind of realization or, rather, are instances of different realizations" (Shapiro 2004, 95).

Adopted and defended also by Polger (2008, 2010, and 2013; Polger and Shapiro 2008; Shapiro and Polger 2012), Shapiro's account has been contested by Aizawa and Gillett (2003, and Aizawa 2009a, 2009b, and 2011), who have offered an alternative account of realization according to which multiple realization would be rather pervasive. For our purposes we do not have to go into the details of this sophisticated and very interesting debate, nor for that matter even go to the fine details of Shapiro's account (though, some of its aspects – e.g., the relativity and intransitivity of the realization relation that it implies – may be relevant for the assessment of, e.g., the strategy of going deeper down science's ontological hierarchy – cf. Polger 2008, 544 n. 5). Suffice it to point out that, applied to the mind-brain relation, this account may undermine the structure-determines-kind assumption. For just as structurally different waiter's corkscrews – a steel one and an aluminum one, say – can realize the function of corkscrewing in exactly the same way, and thus belong to the same realizing type of this function, so may different brain structures realize a given psychological function in the same way and thus belong to the same realizing type of this function. Indeed, based on (1) his account of the realization relation, and (2) the claim that there may well be natural constraints on the kind of structure that is capable of rendering a humanlike psychology, a claim that gets support from instances of neural convergence (i.e., the independent evolution of similar kinds of neural structures), Shapiro argues that (3) it seems plausible that any organ that

exhibits humanlike psychological capacities must also possess various humanlike brain properties (Shapiro 2004, Chaps. 3-4).

However, in taking realizing brain structures to be classified by the way they bring about their realized psychological functions, Shapiro's strategy gives the latter a constitutive role in the classification of the former. So like Bechtel's and Mundale's strategy it also gives a role in the classification of brain structures to the behavioral criteria of psychological functions. Thus, and again like Bechtel's and Mundale's strategy, Shapiro's strategy for defending the One-One assumption must involve a rejection of the reductionist assumption of the mind-brain identity thesis according to which the behavioral criteria of mental kinds do not play a role in determining these kinds. To recapitulate, if the psychological kind of picture naming is identified with a kind of brain structure, and, as by Shapiro's strategy, this brain kind is determined by the behavioral criteria for picture naming, then these behavioral criteria also play a role in determining the mental kind of picture naming.

7.3 A Tint of Functionalism

A *functional property* is a property specified by a job description, or by a certain function this property can perform. Thus, showing the time is a functional property of clocks. According to the functionalist conception of the mind, or *functionalism*, mental kinds are determined by functional properties of the body, which are defined in terms of the role they play as causal intermediaries between perceptual input, other mental states, and behavioral output (Kim 1996, Chap. 5; Antony 2007; Levin 2013). For (an avowedly simplistic) example, a functionalist theory might identify the state of believing that it is raining with the functional property of being in a state that tends to be produced when it is raining, and, given one's belief that by using an umbrella one can avoid getting wet as well as one's desire not to get wet, leads one to take an umbrella. Alternatively, such a theory might identify the state of believing that it is raining with a brain state, all of whose concrete instantiations tend to be produced when it is raining, and, given one's belief that by using an umbrella one can avoid getting wet as well as one's desire not to get wet, causes one to take an umbrella.

Identifying the mental state of believing that it is raining with a higher-level functional property, the first example illustrates the so called *role* version of functionalism (Levin 2013, § 3.4). Identifying the same mental state with a brain state the classificatory criteria of which are constituted by the higher-level functional property that it realizes, the second example illustrates the so called *realizer* version of functionalism (ibid.).

In identifying mental states with brain states that are determined by aspects of higher-level psychological functions and their behavioral classificatory criteria, both the proposal of Bechtel and Mundale and that of Shapiro have significant affinities with realizer functionalism. This is rather ironic, since these proposals seek to defend functionalism's main rival – viz., mind-brain identity theory – by undermining the multiple-realization objection to this rival position.

8. Back to Reverse Inference

The question of whether reverse inference can exceed the limits of behavior based procedures boils down, as we have seen, to the question of the viability of a mind-brain identity thesis which does not give behavior a constitutive role in the classification of the brain kinds with which mental kinds are identified. The question of the viability of the latter thesis depends, in turn, on whether the One-One assumption can be defended against the multiple realization objection. Of the three defense strategies of this assumption that we outlined, the first two – going deeper down science's ontological hierarchy and eliminate and split – are unsuccessful, while the two versions of the third strategy – Bechtel's and Mundale's on the one hand, and Shapiro's on the other hand – may succeed but at the cost of giving behavior a role in the classification of the brain kinds with which mental kinds are identified. This being the case, these two strategies do not yield an account of mind-brain identity capable of grounding a positive answer to the question of whether reverse inference can exceed the limits of behavior based procedures. It follows that philosophical defenses of mind-brain identity may be useless for the purposes of science. Nevertheless, they are not irrelevant to science, since their failure to ground a specific scientific research strategy such as reverse inference or some of its intended uses-cum-goals may form an important negative lesson concerning this strategy.

References

Aguirre, G.K. 2003. "Functional Imaging in Behavioral Neurology and Cognitive Neuropsychology." In *Behavioral Neurology and Cognitive Neuropsychology*, edited by T.E. Feinberg and M.J. Farah, 85–96. New York: McGraw-Hill.

Aizawa, K. 2007. "The Biochemistry of Memory Consolidation: A Model System for the Philosophy of Mind." *Synthese* 155: 65–98.

Aizawa, K. 2009. "Neuroscience and Multiple Realization: A Reply to Bechtel and Mundale." *Synthese* 167: 493–510.

Aizawa, K. and Gillett, C. 2009a. "Levels, Individual Variation, and Massive Multiple Realization in Neurobiology." In *The Oxford Handbook of Philosophy and Neuroscience*, edited by J. Bickle, 539–581. New York: Oxford University Press.

Aizawa, K. and Gillett, C. 2009b. "The (Multiple) Realization of Psychological and Other Properties in the Sciences." *Mind and Language* 24: 181–208.

Aizawa, K. and Gillett, C. 2011. "The Autonomy of Psychology in the Age of Neuroscience." In *Causality in the Sciences*, edited by P.M. Illari, F. Russo, and J. Williamson, 202–223. New York: Oxford University Press.

Allen, C., Grau, J.W., and Meagher, M.W. 2009. "The Lower Bounds of Cognition: What do Spinal Cords Reveal?" In *The Oxford Handbook of Philosophy and Neuroscience*, edited by J. Bickle, 120-142. New York: Oxford University Press.

Anderson, M.L. 2010a. "Neural Reuse: A Fundamental Organizational Principle of the Brain." *Behavioral and Brain Sciences*, 33: 245–266.

Anderson, M.L. 2010b. "Cortex in Context: Response to Commentaries on Neural Reuse." *Behavioral and Brain Sciences* 33: 294–313.

Anderson, M.L., and Penner-Wilger, M. 2013. "Neural Reuse in the Evolution and Development of the Brain: Evidence for Developmental Homology?" *Developmental Psychobiology* 55: 42–51.

Antony, L. 2007. "Everybody Has Got It: A Defense of Non-Reductive Materialism." In *Contemporary Debates in Philosophy of Mind*, edited by B.P. McLaughlin and J. Cohen, 143–159. Oxford: Blackwell.

Bechtel, W. and Mundale, J. 1999. "Multiple Realizability Revisited: Linking Cognitive and Neural States." *Philosophy of Science* 66: 175–207.

Bennett, M.R. and Hacker, P.M.S. 2003. *Philosophical Foundations of Neuroscience*. Oxford: Blackwell.

Bernheim, B.D. 2009. "On the Potential of Neuroeconomics: A Critical (but Hopeful) Appraisal." *American Economic Journal: Microeconomics* 1: 1–41 (NBER Working Paper Series, 13954).

Bickle, J. 2003. *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Dordrecht: Springer.

Bickle, J. 2010. "Has the Last Decade of Challenges to the Multiple Realization Argument Provided Aid and Comfort to Psychoneural Reductionists?" *Synthese* 177: 247–260.

Bourgeois-Gironde, S. 2010. "Is Neuroeconomics Doomed by the Reverse Inference Fallacy?" *Mind & Society* 9: 229–249.

Camerer, C.F., Loewenstein, G., and Prelec, D. 2004. "Neuroeconomics: Why Economics Needs Brains." *Scandinavian Journal of Economics* 106: 555–579.

Camerer, C.F., Loewenstein, G., and Prelec, D. 2005. "Neuroeconomics: How Neuroscience Can Inform Economics." *Journal of Economic Literature* XLIII: 9–64.

Camerer, C.F. 2007. "Neuroeconomics: Using Neuroscience to Make Economic Predictions." *Economic Journal* 117: c26–c42.

Camerer, C.F. 2008a. "Neuroeconomics: Opening the Gray Box." *Neuron* 60: 416–419.

Christoff, K. and Owen, A.M. 2006. "Improving Reverse Neuroimaging Inference: Cognitive Domain Versus Cognitive Complexity." *Trends in Cognitive Sciences* 10: 352–353.

Churchland, P.M. 1982. "Is Thinker a Natural Kind?" *Dialogue* 21: 223–238.

Churchland, P.S. 2008. "The Impact of Neuroscience on Philosophy." *Neuron* 60: 409–411.

Craver, C.F. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Clarendon Press.

Crick, F. 1994. *The Astonishing Hypothesis: The Scientific Search for the Soul*. New York: Touchstone Press.

Dronkers, N.F., Redfern, B.B., and Knight, R.T. 2000. "The Neural Architecture of Language Disorders." In *The New Cognitive Neurosciences*, edited by M. Gazzinga, 949–958. Cambridge, MA: MIT Press.

Edelman, G.M. and Gally, J.A. 2001. "Degeneracy and Complexity in Biological Systems." *Proceedings of the National Academy of Sciences of the USA* 98: 13763–68.

Figdor, C. 2010. "Neuroscience and the Multiple Realization of Cognitive Functions." *Philosophy of Science* 77: 419–456.

Fox, P.T. and Friston, K.J. 2012. "Distributed Processing; Distributed Functions?" *NeuroImage* 61: 407–426.

Gillett, C. 2003. "The Metaphysics of Realization, Multiple Realizability, and the Special Sciences." *The Journal of Philosophy* 100: 591–603.

Greenspan, R.J. 2001. "The Flexible Gnome." *Nature Reviews Genetics* 2: 383–387.

Greenspan, R.J. 2003. "Darwinian Uncertainty." *KronoScope* 3 (2): 217–225.

Harrison, G.W. 2008. "Neuroeconomics: A Rejoinder." *Economics and Philosophy* 24: 533–544.

Henson, R.N. 2005. "What Can Functional Imaging Tell the Experimental Psychologist?" *Quarterly Journal of Experimental Psychology A* 58: 193–233.

Huettel, S.A., Stowe, C.J., Gordon, E.M., Warner, B.T., and Platt, M.L. 2006. "Neural Signatures of Economic Preferences for Risk and Ambiguity." *Neuron* 49: 765–775.

Hutzler, F. 2014. "Reverse Inference is not a Fallacy Per Se: Cognitive Processes Can be Inferred from Functional Imaging Data." *Neuroimage* 84: 1061–1069.

Kable, J.W., and Glimcher, P.W. 2007. "The Neural Correlates of Subjective Value During Intertemporal Choice." *Nature Neuroscience* 10: 1625–1633.

Kenning, P., and Plassmann, H. 2005. "Neuroeconomics: An Overview from an Economic Perspective." *Brain Research Bulletin* 67: 343–354.

Kim, J. 1996. *Philosophy of Mind*. Boulder: WestviewPress.

Kim, J. 2005. *Physicalism, or Something Near Enough*. Princeton: Princeton University Press.

LeDoux, J. 2003. *Synaptic Self: How Our Brains Become Who We Are*. London: Penguin Books.

Lee, V.K. and Harris, L.T. 2013. "How Social Cognition Can Inform Social Decision Making." *Frontiers in Neuroscience* 7: 1–13.

Levin, J. 2013. "Functionalism." *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/functionalism/.

Levin, Y. and Aharon, I. 2011. What's On Your Mind? A Brain Scan Won't Tell." *Review of Philosophy and Psychology* 2: 699–722.

Locke, J. 1975. *An Essay Concerning Human Understanding*. Edited by P.H. Nidditch. Oxford: Clarendon Press.

McClure, S.M., Laibson, D.I., Loewenstein, G., and Cohen, J.D. 2004. "Separate Neural Systems Value Immediate and Delayed Monetary Rewards." *Science* 306: 503–507.

Noppeney, U., Friston, K.J., and Price, C.J. 2004. "Degenerate Neuronal Systems Sustaining Cognitive Functions." *Journal of Anatomy* 205 (6): 433–442.

Noppeney, U., Penny, W.D., Price, C.J., Flandin, G., and Friston, K.J. 2006. "Identification of Degenerate Neuronal Systems Based on Intersubject Variability." *NeuroImage* 30: 885–890.

Owen, A.M. and Coleman, M.R. 2008. "Functional Neuroimaging of the Vegetative State." *Nature Reviews Neuroscience* 9: 235–243.

Page, M.P.A. 2006. "What Can't Functional Neuroimaging Tell the Cognitive Psychologist?" *Cortex* 42: 428–443.

Poldrack, R.A. and Wagner, A.D. 2004. "What Can Neuroimaging Tell Us About the Mind? Insights from Prefrontal Cortex." *Current Directions in Psychological Science* 13: 177–181.

Poldrack, R.A. 2006. "Can Cognitive Processes Be Inferred from Neuroimaging Data?" *Trends in Cognitive Science* 10: 59–63.

Poldrack, R.A. 2008. "The Role of fMRI in Cognitive Neuroscience: Where Do We Stand?" *Current Opinion in Neurobiology* 18: 223–227.

Poldrack, R.A. 2011. "Inferring Mental States from Neuroimaging Data: From Reverse Inference to Large-Scale Decoding." *Neuron* 72: 692–697.

Poldrack, R.A. 2012. "The Future of fMRI in Cognitive Neuroscience." *Neuroimage* 62: 1216–1220.

Polger, T.W. 2008. "Two Confusions Concerning Multiple Realization." *Philosophy of Science* 75: 537–547.

Polger, T.W. 2009. "Evaluating the Evidence for Multiple Realization." *Synthese* 167: 457–472.

Polger, T.W. 2010. "Mechanisms and Explanatory Realization Relations." *Synthese* 177: 193–212.

Polger, T.W. 2011. "Are Sensations Still Brain Processes?" *Philosophical Psychology* 24: 1–21.

Polger, T.W. 2013. "Realization and Multiple Realization, Chicken and Egg." *European Journal of Philosophy* 21 (1): 1–16.

Polger, T.W. and Shapiro, L.A. 2008. "Understanding the Dimensions of Realization." *Journal of Philosophy* 105: 213–222.

Puccetti, R. 1977. "The Great C-Fiber Myth: A Critical Note." *Philosophy of Science* 44: 303–305.

Price, C.J. and Friston, K.J. 2002. "Degeneracy and Cognitive Anatomy." *Trends in Cognitive Sciences* 6 (10): 416–421.

Ramsey, J.D., Hanson, S.J., Hanson, C., Halchenko, Y.O., Poldrack, R.A., and Glymor, C. 2010. "Six Problems for Causal Inference from fMRI." *NeuroImage* 49: 1545–1558.

Richardson, R.C. 2009. "Multiple Realization and Methodological Pluralism." *Synthese* 167: 473–492.

Sanfey, A.G., Loewenstein, G., McClure, S.M., and Cohen, J.D. 2006. "Neuroeconomics: Cross-Currents in Research on Decision-Making." *Trends in Cognitive Sciences* 10: 108–116.

Shapiro, L.A. 2000. "Multiple Realizations." *Journal of Philosophy* 97: 635–654.

Shapiro, L.A. 2004. *The Mind Incarnate*. Cambridge MA: The MIT Press

Shapiro, L.A. 2008. "How to Test Multiple Realization?" *Philosophy of Science* 75: 514–525.

Shapiro, L.A. and Polger, T.W. 2012. "Identity, Variability, and Multiple Realization in the Special Sciences." In S. Gozzano and C.S. Hill (eds.) *New Perspectives on Type Identity: The Mental and the Physical*, 264-287. Cambridge: Cambridge University Press.

Squire, L.R. 2004. "Memory Systems of the Brain: A Brief History and Current Perspective." *Neurobiology of Learning and Memory* 82: 171–177.

Stallen, M. and Sanfey, A.G. 2013. "The Cooperative Brain." *The Neuroeconomist* 19: 292–303.

Sullivan, J.A. 2008. "Memory Consolidation, Multiple Realization, and Modest Reductions." *Philosophy of Science* 75: 501–513.

Tononi, G., Sporns, O., and Edelman, G.M. 1999. "Measures of Degeneracy and Redundancy in Biological Networks." *Proceedings of the National Academy of Sciences of the USA* 96: 3257–3262.

Van Horn, J.D., and Poldrack, R.A. 2009. "Functional MRI at the Crossroads." *International Journal of Psychopisiology* 73: 3–9.

Young, L. and Saxe, R. 2009. "An FMRI Investigation of Spontaneous Mental State Inference for Moral Judgment." *Journal of Cognitive Neuroscience* 21: 1396–1405.

Journal of Cognition and Neuroethics

Informed Consent in Organ Donation and Abandonment of the Dead-Donor Rule

Matthew Phillip Mead
University of Michigan Medical School

Biography

Matthew Mead is currently a fourth year medical student at the University of Michigan Medical School. His interest in medical ethics, particularly in the ethics of organ transplantation and the various definitions of death, began as an undergraduate student at the University of Michigan-Flint where he completed a Bachelor of Science in Biology and Bachelor of Arts in Philosophy. He current research interests have broadened to include clinical research in the field of Orthopaedic Surgery, which he intends to pursue after graduation from medical school.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). September, 2015. Volume 3, Issue 2.

Citation

Mead, Matthew Phillip. 2015. "Informed Consent in Organ Donation and Abandonment of the Dead-Donor Rule." *Journal of Cognition and Neuroethics* 3 (2): 47–56.

Informed Consent in Organ Donation and Abandonment of the Dead-Donor Rule

Matthew Phillip Mead

Abstract

There has been considerable discussion regarding the ethics of organ transplantation and the dead-donor rule (DDR). Much of the medical and philosophical literature reveals inherent difficulties in definitions of death and the appropriate time to begin organ procurement. In this essay, an argument is presented for abandoning the DDR and switching to a practice in which donors are informed of the conditions under which their organs will be removed, rather than the current practice of requiring a declaration of death. Informed organ donation consent (IODC) would allow for greater transparency in the organ procurement process and alleviate many of the ethical concerns raised in the literature today surrounding these practices. This has the potential to improve public trust of organ procurement and increase the numbers of donors.

Keywords

Dead Donor Rule, Death, Definition/Determination of Death, Donation/Procurement of Organs/Tissues, Vital Organ Donation

In recent years, there has been considerable debate surrounding the ethics of organ transplantation in both the medical and philosophical literature. Some of this debate has been focused on organ procurement practices and the criteria used to justify the appropriate time to initiate organ procurement. Properly defining the conditions under which a donor's organs can be removed is particularly important, as it represents an important safeguard against violating individual rights. Early in this discussion, it was recognized that vulnerable donor groups needed protection, such as the poor, the elderly, prisoners, the mentally handicapped, and patients who it was unclear whether they were alive or dead. From this came one of the most fundamental principles in the protection of organ donors: the requirement that a potential donor be declared dead prior to the removal of their organs, known thereafter as the dead donor rule (DDR) (Robertson 1999). However, as will be argued, the DDR unnecessarily complicates organ procurement and fails to provide consistent conditions under which a person's organs will be removed as a result of there being multiple definitions of death. Because of this, the DDR should be abandoned and replaced with a practice in which donors are informed of the conditions under which their organs will be removed. This "informed organ donation consent" (IODC) would allow for greater transparency in the organ procurement process

and alleviate many of the ethical concerns raised in the literature today surrounding these practices.

To better understand this, it is useful to explore the reasons for abandoning the DDR. At first glance, the DDR appears to provide a simple and reliable way in which donors can be protected from abuses. However, the DDR does nothing to define what it actually means for a donor to be dead. Medical and philosophical literature reveals that death is not as clear of a concept as once thought. For example, one might wonder whether a person in an irreversible coma could be dead, or whether a person whose heart and lungs function only as a result of mechanical ventilation could actually be alive. Under the DDR, correctly answering these questions is of great importance, as it dictates whether organ procurement can begin or not.

In an effort to clarify the concept of death, three main definitions of death have been developed, each of which has been used in the DDR. The most traditional of these is the cardiopulmonary definition of death (CPD), which is usually defined as the irreversible cessation of heart and lung function (Iltis & Cherry 2010; Kerridge *et al.* 2002). This is in keeping with the customary notion of death, where breathing and pulses are signs of life and the absence of these are a sign of death. However, with the advent of mechanical ventilation and electronic defibrillation, many have concluded the CPD to be inadequate. One criticism of the CPD is that a patient who lacks brain and brainstem activity would be still be considered alive while on a mechanical respirator. Essentially, patients who would die otherwise can be kept alive under the CPD for months or even years as long as these cardiopulmonary supports are in place, even though they lack the type of brain function that many people feel is important for life.

Recognizing these inadequacies, a committee at Harvard Medical School released a report in 1968 which redefined the concept of death in humans into what is now called the brain-death definition (BDD) (Harvard 1968). According to the BDD, death is defined as the irreversible cessation of function of the brain and brainstem. This was later adopted as the primary legal definition of death by a 1981 Presidential Committee (President's Commission 1981). A variety of tests have been developed to determine whether a patient's brain and brainstem are functioning. These tests classify a patient as dead if they show a lack of awareness to external stimuli and unresponsiveness to painful stimuli, a lack of spontaneous muscular movement and respirations, and lack of key reflexes. Such findings include fixed, dilated pupils, a lack of eye movement even when the eyes are hit, moved, or stimulated by cold water in the ear (caloric reflex test), and a lack of response to noxious stimuli. This definition is useful because it helps to declare death in unclear situations, such as when a patient is on a mechanical ventilator.

Despite its strengths, several objections have been presented against the BDD. The first is that individuals can fulfill all of the diagnostic tests for determination of brain death, but still retain evidence of integrated brain function at the mid-brain and brainstem level, and in some cases may even continue to have some evidence of cortical function (Chiong 2005). For example, many different hormones (such as growth hormone, prolactin, thyroid stimulating hormone, cortisol, and vasopressin) that are regulated in the brain continue to be regulated after determination of brain death (Schrader *et al.* 1980). Additionally, a flat electroencephalogram is not always observed in patients who fulfill the diagnostic requirements for brain death. One study has found that 20% of patients who fulfill the diagnostic tests for brain death still show some level of electrical activity on an electroencephalogram (Grigg *et al.* 1987). Other clinicians have observed that patients who have fulfill the diagnostic tests for brain death often respond to pain with significant increases in both heart rate and blood pressure (Shewmon 1998; Wetzel *et al.* 1985).

While none of these are criticisms of the BDD itself (but rather criticisms of the inadequacy of the tests used to diagnose brain death), they demonstrate that despite considerable study, properly defining conditions under which the criteria for the BDD have been met remains elusive. Robert Truog, an early proponent of the abandonment of the DDR writes,

> This evidence points to the conclusion that there is a significant disparity between the standard tests used to make the diagnosis of brain death and the criterion these tests are purported to fulfill. Faced with these facts, even supporters of the current statuses acknowledge that the criterion of 'whole-brain' death is only an 'approximation.' (Truog 1997)

A third definition of death has also been proposed, known as the higher-brain function definition (HBF). The HBF definition holds that it is the potential for consciousness which differentiates between life and death. According to the HBF definition, death is the irreversible loss of personhood. If we assume consciousness is necessary for personhood, then the irreversible loss of consciousness represents the death of the person. Thus, under the HBF definition, patients in irreversible comas, newborns with anencephaly, and patients in persistent vegetative states (PVS) are all considered dead. It is with PVS patients that many find difficulty, as these patients retain complete or partial hypothalamic and brainstem functions, such as thermoregulation and the ability to swallow (Monti *et al.* 2010). They occasionally smile, cry, grunt, or moan in

response to internal stimuli. In contrast to CPD and BDD patients, irreversible coma and PVS patients do not require life-sustaining machinery (other than feeding tubes), as the areas of the brain that control respiration, hormone levels, blood pressure, heart rate, and gastrointestinal function remain intact.

Not surprisingly, a major criticism of HBF is that it classifies these PVS patients as dead. Many intuitively feel that these patients are alive, perhaps even "more" alive than patients who require mechanical respirators. Similar to a BDD patient, a PVS patient continues to grow and mature sexually. They can become ill and fight off infections. They maintain homeostatic control of hormone regulation, body temperature, and fluid balance (Schrader *et al*. 1980). Patients in PVS retain a gag reflex, exhibit evidence of normal sleep cycles, and periodically yawn. Their eyes will track light and in some cases even moan when their muscles are overly stretched. For many, it is difficult to conclude that PVS patients are dead. This stems from the difficulty in separating the biologic processes that are associated with life from the death of the person. According to the HBF definition, these patients lack the potential for consciousness and have lost personhood, and therefore are dead. Critics, such as David DeGrazia, have argued against this view on the grounds that we, as humans, are not *essentially* persons (Degrazia 1999, 2002, 2006). In other words, there are periods in our lives in which we exist as nonpersons, such as during infancy or severe dementia. Yet during these times, we do not consider ourselves as being dead. Thus, the status of one's personhood does not dictate whether a person is alive or dead and thus, neither would the irreversible loss of consciousness.

The purpose of this discussion has not been to promote one stance over another, but rather to show the significant disagreements among the academic community concerning the topic of death. These disagreements have resulted in numerous ethical dilemmas involving the DDR and organ procurement, where physicians must turn to one theory of death over another in order to justify the initiation of organ procurement. An excellent example of this is what has been termed donation after cardiac death (DCD). This practice, first developed at the University of Pittsburg Medical Center in 1992, was a novel method in obtaining organs from patients who were (1) expected to die shortly, (2) demonstrated a wish to donate, and (3) who had verified do not resuscitate order on record (Pittsburg 1993). Although specific procedures have changed since its inception, the overall process remains similar. At some point in the treatment, the decision is made by family or medical personnel that it is appropriate to remove life support from these patients because they show no hope for improvement. However, prior to the removal of life support, these patients are taken to surgery and prepped for organ procurement. Additionally, prospective organ recipients and their respective surgery teams are informed

of the organs which would be available shortly. This allows adequate time for these teams to organize and prepare the prospective organ recipient for surgery as well. Then, when the time is appropriate, life-sustaining interventions are removed from the donor and these patients are allowed to "die" under CPD criteria. Following pronouncement of death, the donor is place back on life-support to maintain adequate blood flow to the organs to be removed. These DCD procedures, allow the patient to be declared legally dead prior to organ procurement and allow for greater quality of transplantable organs by reducing the amount of ischemic damage.

There has been considerable debate over the ethics of these practices. Some have argued that DCD procedures devalue appropriate end-of-life care.[1] Others have expressed concern over the appropriateness of administering anticoagulants and vasodilator medications to these patients prior to death, as they are given to enhance the viability of the transplantable organs but can be detrimental to the prognosis of the not-yet-dead patient (Menikoff 2002). There has also been concern about the length of time patients should remain asystole. The original University of Pittsburg protocol required patients to remain asystole for two minutes prior to declaring death. However, there have been patients that returned to cardiac rhythms after more than two minutes of asystole (Adhiyaman *et al.* 2007; Rady *et al.* 2007). To palliate this concern, some institutions have increased the asystole period from two minutes to five minutes. However, peer reviewed medical literature demonstrates some patients returning to cardiac rhythms after more than ten minutes (Adhiyaman *et al.* 2007; Hornby *et al.* 2010). The most serious criticism, however, has been over whether DCD donors could ever actually be declared dead under CPD criteria, since this requires the *irreversible cessation* of respiration and circulation. In DCD procedures, it is clearly evident that irreversibility has not been met if the goal is to restart the heart and lungs after the declaration of death.

The purpose in describing DCD procedures and their criticisms has been to demonstrate the problems that the DDR creates for organ procurement. Physicians are forced to jump through hoops in order abide by the DDR and preserve the intentions of the donor. The easiest solution to this problem is to eliminate the DDR altogether and replace it with an organ donation process by which donors are simply informed of the conditions under which their organs will be removed, a practice I term as informed organ donation consent (IODC). These conditions would not require that a person be declared dead (although they may or may not be, depending on their specific circumstances and

1. For a short commentary on these concerns, see Rady, M.Y., J.L. Verheijde, and J. McGregor. 2006. "Organ Donation after Circulatory Death: The Forgotten Donor?" *Crit Care* 10 (5) 166: 1–3.

what definition of death you subscribe to). Highlighting the unnecessary connection between organ donation and death, Robert Truog writes,

> That patients be dead before their organs are recovered is not a foundational ethical requirement. Rather, by blocking reasonable requests from patients and families to donate, the DDR both infringes donor autonomy and unnecessarily limits the number and quality of transplantable organs. (Truog *et al.* 2013)

At this point, one might ask what the criteria would be in IODC during which organ procurement could begin. In my view, the conditions could be the same clinical criteria used for defining brain death today. That is, a patient might consent to having their organs removed if they arrived at a condition in which they lacked spontaneous respirations, had fixed, dilated pupils, a lack of eye movement when their eyes were hit, moved, or stimulated by ice cold water in the ear, etc. Thus, the transition from the DDR to IODC would not require a significant change in practice. Much of the clinical criteria used today to declare death could very well be used to declare a patient suitable for organ procurement. Others could certainly argue for different clinical criteria that they find suitable. The key distinction though, is that whether or not these patients were dead from a legal or philosophical perspective would be irrelevant.

With IODC in place, organ procurement practices could be standardized or it could be up to the prospective donor. This means that patients could define the conditions on their own (similar to a living will or a DNR order) or one set of conditions (such as the BDD, CPD, or HBF criteria) could be applied universally. Importantly, the elimination of the DDR and the transition to IODC eliminates the logical problems and societal misconception about defining death. Because of this, IODC allows for greater accuracy in determining the appropriate time for procurement to begin and better protects donor autonomy.

One concern with IODC may be that in eliminating the requirement for a legal declaration of death prior to organ procurement, it is possible that the practice of organ transplantation might lose public trust and support. This would have the potential to decrease the number of available transplantable organs. However, empirical evidence supports that the DDR is not essential for public trust in organ donation. In a 2003 study, researchers at Case Western Reserve University examined factors related to families' understanding of brain death and how those factors affected decisions about organ donation (Siminoff *et al.* 2003). Their results indicated that in a sample of over four-hundred families who had family members who had been declared dead using BDD

diagnostic tests, only 28% could give a correct definition of BD. Furthermore, they found that there was no association between a willingness to donate and having an accurate understanding of brain death.

Another concern with IODC might be that the elimination of the DDR in organ procurement would lead to a slippery slope, where organs are removed from patients who have reasonable chances of survival. This problem is alleviated if patients are required to choose organ procurement conditions in which they are very close to death and show no hope of recovery. As mentioned previously, these would likely be the criteria already used to determine brain death, cardiopulmonary death, or HBF. Thus, there would be little real change in practice, other than the elimination of the need for practices such as DCD.

In conclusion, the benefits of standardizing organ donation policies, as well as freeing practitioners from relying on such problematic concepts as "legally dead" are major advantages of abandoning the DDR. IODC would provide practitioners with the opportunity to educate members of the public on the complexities of the processes of death. Patients and their family members would then be able to make truly informed decisions regarding organ donation. This has the potential to foster public support for organ donation and begin to address the severe organ shortages which greatly limit transplantation today.

References

Adhiyaman, V., Adhiyaman, S., & Sundaram, R. 2007. "The Lazarus Phenomenon." *Journal of the Royal Society of Medicine* 100 (12): 552–557.

Chiong, W. 2005. "Brain Death without Definitions." *Hastings Center Report* 35 (6): 20–30.

Degrazia, D. 1999. "Advance Directives, Dementia, and 'The Someone Else Problem.'" *Bioethics* 13 (5): 373–391.

Degrazia, D. 2002. "Are We Essentially Persons? Olson, Baker, and a Reply." *The Philosophical Forum* 33 (1): 101–120.

DeGrazia, D. 2006. "Moral Status, Human Identity, and Early Embryos: A Critique of the President's Approach." *The Journal of Law, Medicine & Ethics* 34 (1): 49–57.

Grigg, M. M., Kelly, M. A., Celesia, G. G., Ghobrial, M. W., & Ross, E. R. 1987. "ELectroencephalographic Activity after Brain Death." *Archives of Neurology* 44 (9): 948–954.

Hornby, K., Hornby, L., & Shemie, S. D. 2010. "A Systematic Review of Autoresuscitation after Cardiac Arrest." *Crit Care Med* 38 (5): 1246–1253.

Committee of Non-Heart-Beating Transplantion. 2000. "The Scientific and Ethical Basis for Practice and Protocols Division of Health Care Services." *Institute of Medicine*.

Iltis, A. S., & Cherry, M. J. 2010. "Death Revisited: Rethinking Death and the Dead Donor Rule." *Journal of Medicine and Philosophy* 35 (3): 223–241.

Kerridge, I. H., Saul, P., Lowe, M., McPhee, J., & Williams, D. 2002. "Death, Dying and Donation: Organ Transplantation and the Diagnosis of Death." *Journal of Medical Ethics* 28(2): 89–94.

Menikoff, J. 2002. "The Importance of Being Dead: Non-heart-beating Organ Donation." *Issues in Law and Medicine* 18 (1): 3–20.

Monti, M. M., Laureys, S., & Owen, A. M. 2010. "The Vegetative State." *BMJ* 341: c3765.

Rady, M. Y., Verheijde, J. L., & McGregor, J. 2007. "'Non-heart-beating,' or 'cardiac death,' Organ Donation: Why We Should Care." *Journal of Hospital Medicine* 2 (5): 324–334.

Rady, M.Y., J.L. Verheijde, and J. McGregor. 2006. "Organ Donation after Circulatory Death: The Forgotten Donor?" *Crit Care* 10 (5) 166: 1–3.

President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavior Research. 1981. "Defining Death: Medical, Legal and Ethical Issues in the Deter-mination of Death."

Robertson, J. A. 1999. "Delimiting the Donor: The Dead Donor Rule." *Hastings Center Report* 29 (6): 6–14.

Ad Hoc Committee of the Harvard Medical School. 1968. "A Definition of Irreversible Coma-Report of the Ad Hoc Committee of the Harvard Medical School to Examine the Definition of Brain Death." *Journal of the American Medical Association* 205 (6): 337–340.

Schrader, H., Krogness, K., Aakvaag, A., Sortland, O., & Purvis, K. 1980. "Changes of Pituitary Hormones in Brain Death." *Acta Neurochiurgica* 52 (3–4): 239–248.

Shewmon, D. A. 1998. "'Brainstem Death,' 'Brain Death' and Death: A Critical Re-evaluation of the Purported Equivalence." *Issues in Law and Medicine* 14 (2): 125–145.

Siminoff, L. A., Mercer, M. B., & Arnold, R. 2003. "Families' Understanding of Brain Death." *Progress in Transplantation* 13 (3): 218–224.

Truog, R. D. 1997. "Is It Time to Abandon Brain Death?" *The Hastings Center Report* 27 (1): 29–37.

Truog, R. D., Miller, F. G., & Halpern, S. D. 2013. "The Dead-Donor Rule and the Future of Organ Donation." *New England Journal of Medicine* 369 (14): 1287–1289.

University of Pittsburgh Medical Center Policy and Procedure Manual. 1993. "Management of Terminally Ill Patients who May Become Organ Donors after Death." *The Kennedy Institute of Ethics Journal* 3(2): A1–15.

Wetzel, R. C., Setzer, N., Stiff, J. L., & Rogers, M. C. 1985. "Hemodynamic Responses in Brain Dead Organ Donor Patients." *Anesthesia & Analgesia* 64 (2): 125–128.

Journal of Cognition and Neuroethics

More Than Meets the fMRI: The Unethical Apotheosis of Neuroimages

Eran Shifferman
Edmond J. Safra Center for Ethics, Tel Aviv University

Biography

Eran Shifferman is an independent researcher focusing on various aspects of the evolution of cognition. He began his academic journey as an ethologist and then switched to philosophy of biology and psychology, and finally ethics and science communication. In his work Shifferman attempts to provide broad palate accounts of the emergence, persistence, propagation and fixation of cognitive novelties along the phylogenetic tree. While doing so, Shifferman is critical of the many discourses he's using and tries to make these scientific products more accessible to the public.

Acknowledgments

I'd like to deeply thank Snait Gissis and Aida Robles-Gomez for attentive and insightful readings; the fellows at Edmond J. Safra Center for Ethics in Tel Aviv University for their helpful comments; and Christine Schwabb and Gila Blitz for their comments

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). September, 2015. Volume 3, Issue 2.

Citation

Shifferman, Eran. 2015. "More Than Meets the fMRI: The Unethical Apotheosis of Neuroimages." *Journal of Cognition and Neuroethics* 3 (2): 57–116.

More Than Meets the fMRI: The Unethical Apotheosis of Neuroimages

Eran Shifferman

Abstract

The following is an attempt at a multifaceted critique of cognitive neuroscience's use of images born of blood-oxygen-level dependent functional magnetic resonance imaging (BOLD-fMRI) in support of nesting cognitive functions in specific brain regions. It is an exploration of problems associated with three levels of producing functional neuroimages (NIs): the technological, the methodological, and the philosophical. My goal is not merely to map the spectrum of problems associated with the use of BOLD-fMRI NIs use in cognitive neuroscience. Rather, it is to use this map to support the claim that functional neuroimaging all too often amounts to unethical science, one where the generators of data overlook significant shortcomings of their tools of the trade and press forward with producing claims about the nature of the mind-brain link, which are too strong to be supported, by exploiting the strong appeal of their meticulously crafted images. These claims filter through to have a significant impact on the oblivious public over cardinal topics in psychology and philosophy such as behavior, emotions, consciousness, cognition, and the self.

Keywords

Neuroimaging, Ethics, Cognitive neuroscience, Reductionism, Evolution, Philosophy of Biology, fMRI, BOLD, Essentialism, Public Understanding of Science, Representation, Aesthetics

Introduction

Functional neuroimages (NIs) are vivid, colorful renditions of the brain based on raw numerical data generated by MRI scans taken as subjects perform a cognitive task, and their ensuing extensive statistical manipulations. Progressively, NIs have become a mainstay of neuroscientific arguments concerning the neural underpinnings of cognition and behavior, and have crossed over to the mass media as a canonical representation of the brain. Within the discipline of cognitive neuroscience, NIs are used to support the argument that (at best) particular networks or (more commonly) individual brain regions house specific cognitive, behavioral or emotional phenomena. The long list of cognitive and emotional traits that functional neuroimaging studies attempt to map onto the brain includes (but not limited to): belief (Harris, Sheth, and Cohen 2008; Harris et al. 2009; Kapogiannis et al. 2009; Neubauer 2014; Beauregard and Paquette 2006); humor (Bartolo et al. 2006; Chan et al. 2012; Sawahata et al. 2013); political orientation (Schreiber et al. 2013; Ahn et al. 2014); love (Bartels and Zeki 2000, 2004; Aron et al.

2005; Wan et al. 2014; Fusar-Poli and Broome 2007); moral behavior (Cikara et al. 2014; Yoder and Decety 2014); deception (Aharoni et al. 2013; Koster-Hale et al. 2013; Yang et al. 2014); happiness (Kong et al. 2015); "cultural" differences (Han and Ma 2014); and even stock market forecast (Smith et al. 2014) and response to reality TV shows (Melchers et al. 2015). In their review of NI studies, Gabrieli et al. (Gabrieli, Ghosh, and Whitfield-Gabrieli 2015) boldly claimed that the ability to predict future behavior using NIs is "a humanitarian and pragmatic contribution of human cognitive neuroscience to society".

This paper brings together for the first time four disparate lines of criticism of various aspects of BOLD-fMRI research within cognitive neuroscience that dampen the stronger neurocognitive arguments concerning the neural basis of the mind. The technological criticism has two components to it: on the one hand is the physiological nature of the BOLD signal and its relation to neural activity, and on the other hand are the limitations and parameters of the MR machine itself. The second critique is methodological, dealing with the experimental designs and methods as well as statistical tools used to produce NIs, and the impact of the vast array of available design possibilities on them. The third line of investigation is the philosophical, addressing the basic tenets and widely accepted (and practiced) underlying assumptions concerning the link between neuronal activity - as portrayed by functional Nis - and claims about the nature of brain operation, the mind, and the self. Fourthly, the resultant criticism will then be merged with some sociotechnological insights concerning the concept of representation and the power dynamics between scientists and non-professionals. This is done in an attempt to complete and support the main argument of this paper: that the use of fMRI NIs in order to triangulate brain coordinates of consciousness and cognition is direly problematic and yet marketed to non-professionals as bona fide scientific truth that in turn shapes specious public perception.

The Physiological Basis of fMRI

To begin with, let us clarify what BOLD-fMRI allows for. Nerve cells' activation is an energy consuming process that relies on the metabolism of glucose. These metabolic demands depend on blood flow and cause brain blood oxidation status changes. Oxygenated and deoxygenated blood have different magnetic properties, and to determine whether a brain region has been activated in response to a specific stimulus, one scans for an increase in oxygenated blood in that region as a function of conditions that change over time. Thus - crucially - the BOLD signal is an indirect indicator of neural

activity, in which blood and metabolism are the mediators. Cardinal to the validity of the BOLD signal as an indicator of neural activity is neurovascular coupling, a metabolic hypothesis of a one-to-one correspondence between hemodynamic changes (the BOLD signal) and neuronal activity. However, evidence shows exceptions and variations to this assumption, thus complicating the interpretation of fMRI data when analyzing cognition and behavior (Caesar, Thomsen, and Lauritzen 2003; Devor et al. 2008; Sirotin and Das 2009; Ekstron 2010; Jukovskaya et al. 2011; Mishra et al. 2011; Hermes et al. 2012; Siero et al. 2013; Huo, Smith, and Drew 2014; Mayhew et al. 2014). This has led some scientists to warn against assigning too great an importance to fMRI reports (Page 2006; Rossier 2009; Devonshire et al. 2012; Singh 2012).

BOLD Problems

The reliability of the BOLD signal as an indicator of neural activity is cast deeper under shadow as there is more to brain metabolism than just neurovascular coupling. First, firing neurons are not the only cells to cause hemodynamic changes: we can add neurons at subthreshold levels of activation, neurons with varying levels of simultaneous excitation and inhibition, and feedback from local and distant sites (Nair 2005). Non-neuronal entities such as astrocytes and vascular cells also impact brain metabolism (Iadecola 2004; Bélanger, Allaman, and Magistretti 2011; Figley and Stroman 2011; Escartin and Rouach 2013). Second, there is a broad range of different kinds of neurons, each with different genetic makeup, neurotransmitter composition, myelination profiles, spatial structure, connectivity topography, and spatiotemporal functionality, all shaping neuronal operation (Fishell and Heintz 2013) and inflicting unique metabolic changes (Logothetis 2008). Third, the BOLD signal primarily measures the input and processing of neural information within a region and not the output signal transmitted to other brain regions (Logothetis 2003). Without an output, a given brain region is highly unlikely to be generating behavior. Moreover, input to an area, and processing within it, are necessary for the disengagement of that region's function, causing such disengagement to also contribute to the BOLD signal (Page 2006). Fourth, BOLD contrast is generally most intense in the veins downstream from the neural circuits that create metabolic demand, causing the location of a signal change and the location of the presynaptic neural activity to not necessarily correlate (Tancredi and Brodie 2007).

Taken together, the composition and characteristics of local neural activity impede BOLD-fMRI's ability to differentiate between function-specific processing and other neural artifacts; between bottom-up and top-down signals; and between excitation

and inhibition. Also, the magnitude of the signal does not necessarily correlate with the importance of the respective region for the task of interest and cannot be standardized to quantify differences between brain regions, or between tasks within the same region (Huber 2009). It follows that a brain region, a neuroimage, and a cognitive function do not necessarily have a linear one-to-one correspondence. Rather, the relationship is best described as a statistical correlation (Page 2006; Tancredi and Brodie 2007; Logothetis 2008). Such a correlation does not suffice to justify a strong localization argument.

BOLD-fMRI - Technological issues

Exploring the nature of the BOLD signal allowed us to lay the foundation for discussing the first critique, that of the MR equipment itself. In this section we discuss technological hurdles put in front of cognitive neuroscientists when attempting to draw conclusions from BOLD-fMRI data to cognition and behavior[1].

Spatiotemporal Limitations

There are two ways by which spatiotemporal dissonance between the actual neural activity and the ensuing BOLD signal may arise. The first is caused by apparatus limitations. A temporal limitation is levied due to the fact that there is a discord between the time required to generate a given cognitive function (several tens of milliseconds) and that required for an MR machine to collect enough raw data (2-6 seconds) (Haxby, Courtney, and Clark 1998). A spatial limitation is levied since MR has a resolution of about 0.1 mm, while single cell activity operates at three to four orders of magnitude smaller (Hardcastle and Stewart 2002)[2]. The second cause of dissonance is the fact that the BOLD signal is not a direct measure of neural activity, and thus the signal may be temporally and even spatially out of register with the activity changes that are ultimately the phenomenon of interest. In order to bridge this gap we have to make a variety of assumptions about the temporal and spatial relationships between blood flow changes and neural activity, the validity of which is at times questionable (Roskies 2008). Another

1. Logothetis' (Logothetis 2008) excellent discussion of the plethora of shortcomings of current use of BOLD-fMRI in neurocognitive sciences should serve as reference for the many technical details this paper cannot fully explore.

2. New advancements allow for a significantly improved resolution and sample rate, yet these are not widespread and thus not enough to observe complex spatial-temporal orchestration of brain activity that underlies cognition (<10 msec) (Buckner 2003). Therefore, what cortical processes exactly the BOLD signal does and does not represent is still far from clear (Goense, Whittingstall, and Logothetis 2012).

issue concerns brain anatomy: areas adjacent to sinuses (e.g. the orbital frontal cortexes) hold air inside, and this proximity may distort the image (Kringelbach and Rolls 2004), particularly with increasing magnetic field strength (Devlin et al. 2000).

Resolution Problems

A typical MRI voxel contains 5.5 million neurons, ~$4*10^{10}$ synapses, 22 km of dendrites and 220 km of axons (Logothetis 2008). These astronomical numbers include a wide array of neuronal types, which necessarily forces heterogeneity upon any given region of interest. In addition, "flexible" neurons can represent different abstract rules or categories in a temporal and context-dependent manner (Duncan 2001), thus making the definition of the functional role of brain regions a cumbersome task. Furthermore, all the processing within a given voxel involves an extensive range of inputs from other brain regions (Hardcastle and Stewart 2002). Consequentially, conclusions concerning the neuronal activity of a given voxel are governed by the haphazard content of a given sample, thus underestimating when neurons actually respond and under what conditions (Hardcastle and Stewart 2002; Roskies 2008; Meinertzhagen et al. 2009).

One way of tackling this problem is via improved resolution, which allows for more accurate reading of the hemodynamic metabolic changes. Accuracy is measured by the signal-to-noise ratio (higher ratios indicate better quality). The strength of the MR machine's primary magnetic field is of paramount importance to the signal-to-noise ratio, such that stronger fields allow for greater acquisition of significant voxels (Hoenig, Kuhl, and Scheef 2005; Alvarez-Linera; García-Eulate et al. 2011; Wardlaw et al. 2011). Most MRI machines used in neurocognitive science operate at 1.5-3 Tesla (Logothetis 2008; Regatte 2014). However, not only are identified areas likely to morph and grow/shrink; some are likely to be statistically significantly active (if not maximally so) when investigated under higher magnet power, even if this activity is below threshold at 1.5 T (Figdor 2010). So it follows that at least some regions of activation established by using low-frequency scanners may actually be artifacts of their magnet power. Unfortunately, tinkering with any parameter, be it magnetic field strength or voxel size, also impacts negatively on the signal-to-noise ratio, and a delicate balancing game is required, all depending on the research question and study limitations. The reality is that there is a myriad of variables that have a direct and immediate impact on the final NI, and there is very little standardization that allows minimization between- (and even within-) lab procedural variability (Bennett and Miller 2010).

Admittedly, technological issues are being addressed[3]. However, the persistent problem, in the context of this paper, is the fact that the data already generated with older technology is not scrutinized and re-evaluated with improved technology (Vul and Pashler 2012), thus passively allowing possible erroneous data to solidify as a valid reference. Therefore, considering MR technology alone, we have ample reason to doubt strong localization arguments resting on such NIs (see the following discussion of replication, page 8).

BOLD-fMRI - Methodological Impediments

As stated earlier, neuroimaging concerns extend beyond technological facets. This section presents the most pertinent methodological tools used to spawn NIs and the difficulties associated with them: localization, raw data processing, statistical manipulation, variability, and standardization.

Localization

Localization is the attempt to nest a specific cognitive/emotional attribute in a specific brain region. Localization is a central goal of cognitive neuroscience and it is deeply rooted in the history of neuroscience, long before any mapping technology was available (Naneix 2009). Localization has waxed and waned, but the arrival of modern imaging techniques, primarily that of MRI, has brought it to new heights (Vul and Pashler 2012; Klein 2012). Importantly, with the exception of extensively studied peripheral cognitive areas, few neuroimaging localization attempts are beyond controversy (Figdor 2010; Ihnen et al. 2009; Ball et al. 2009). Here, I investigate the localization project by looking at its components and highlight major concerns associated with it.

Subtraction

Localization via BOLD-fMRI is made possible by subtraction. Subtraction entails imaging subjects performing a mixed sequence of two different tasks that are (supposedly) separated by a single cognitive element, ending up with two different time series that can be compared to verify whether the activity in the region of interest was different between the two tasks. Once performed, the image of the "simpler" task is subtracted from the more complex one, creating a difference image that (ideally) has

3. E.g., cutting edge fMRI technology today allows for a single-cell resolution with 17.2T (Radecki et al. 2014). Currently, even 7T is not FDA-approved, though it's popularity rapidly increases (Kraff et al. 2015).

isolated an area of increased or decreased activation. That area is considered to be the seat of the additional cognitive element separating the two tasks (Vul et al. 2009). Subtraction relies on the pure insertion assumption: the amount of additional activity attributable to the interaction between the new and the old tasks is zero. Subtraction hinges on several strong tenets: a) cognitive processing is highly modular; b) the brain is a serial processor; c) cognitive functions are linearly additive so there are no qualitative changes upstream on the shared components of experimental and control tasks and can therefore be subtracted from one another; d) each task invokes a minimum set of components for successful performance (van Orden and Paap 1997; Peterson 2003). At the neural level, the assumption is that the difference in neural activity during baseline and task is due entirely to the new task and does not represent any influence on or interaction with the baseline activity. This assumption ignores the possibility that the additional neural activity may be neither sufficient nor necessary for the presumed-to-be purely additional task (Figdor 2010).

Furthermore, as demonstrated earlier, due to the nature of the BOLD signal and the technological limitations of MR sampling, subtraction cannot determine whether the differences in activity are due to the cognitive process assigned to them a posteriori or due to something else occurring concurrently but coincidentally. Ironically, the lower the sensitivity of the MR apparatus, the better it is for localization: low signal-to-noise ratios allow for only a few statistically significant differences across conditions to be found (Hardcastle and Stewart 2002). Also, pure insertion does not fit well with observations indicating that neural processing utilizes spatiotemporal feedback connections between multiple regions (Poldrack 2010), including even partial information transfer between stages of processing (Miller and Hackley 1992; Bichot, Rao, and Schall 2001). If the pure insertion assumption fails, then there is no way to determine what cognitive processes are reflected in the activation observed in the subtraction experiment.[4]

Reverse Inference

Even if cognitive neuroscientists were to dispense with the problems associated with pure insertion, they would still have to address seriously another major concern of localization, that of reverse inference: inferring the operation of a specific cognitive trait

4. These criticisms have not gone unnoticed, and cognitive neuroscientists have invested effort in adopting designs that attempt to circumvent pure insertion (Poldrack 2010; Kawabata Duncan and Devlin 2011). However, Poldrack (Poldrack 2010) argues that the new methods fail to establish a link between task characteristics and the cognitive processes they are supposed to represent.

based on activation of a specific brain region, reasoning backwards from activation to cognitive operation. For example, activation in the amygdala is interpreted as reflecting fear or negative emotion, even though it can be equally active for positive outcomes (Bunzl, Hanson, and Poldrack 2010). Several observations cast a shadow over reverse inference. First, pluripotency[5] virtually negates reverse inference because activity of a region in response to two or more contexts cannot be used as evidence assigning that region exclusively to only one of them (Haxby et al. 2001; Henson 2005; Price and Friston 2005). Second, evidence shows that the response of regions *other* than those responding maximally to a given stimulus could also predict which stimulus was presented (Haxby et al. 2001; Mole et al. 2007). Together, this evidence shows that reverse inference reflects the logical fallacy of affirming the consequents (Poldrack 2006; Klein 2012).

Additional Concerns over Localization:

Going beyond the problematic pure insertion and reverse inference, other shortcomings of localization have been highlighted. First, traditionally, fMRI studies perform only a handful of scans of many subjects over relatively short time period. Contrary to that, Gonzalez-Castillo et al. (Gonzalez-Castillo et al. 2012) scanned very few subjects for a long cumulative duration, over many scanning sessions. Their analysis shows that 70%-90% of all voxels were labeled as active, suggesting that localization is made possible due to insufficient statistical power. If, as Gonzalez-Castillo and colleagues (Gonzalez-Castillo et al. 2012) suggest, the entire brain is involved in even minor tasks, then the dichotomy between active and inactive regions is of no scientific relevance (Stelzer et al. 2014), and the localization project is severely weakened.

Secondly, early meta-analyses have shown that localization is elusive since the execution of cognitive functions relies on networks connecting many regions with a high degree of spatiotemporal variability, inter-individual differences and context-dependence (McIntosh 2000; Cabeza and Nyberg 2000; Phan et al. 2002; Gerlach 2007; Buchsbaum and D'Esposito 2008; Dolcos, Denkova, and Dolcos 2013).

Cumulatively, the idea of a one-to-one mapping of cortical activation to high-level cognitive processes suggested by NIs seems like an oversimplification of a more complex

5. Pluripotency is the ability of the same neural entity to perform (or cause or be involved in the performance of) several tasks (Nair 2005; Price and Friston 2005; Henny et al. 2012; Lee, Soares, and Beique 2012). Pluripotency connects with the philosophical concept of multiple realizability (the claim that the same mental attribute can be generated by more than one physical substrate) and to the biological concept of degeneracy (the ability of structurally different elements to yield the same output; (Edelman and Gally 2001)).

many-to-many mapping (Just and Varma 2007). This is exemplified by a letter sent to *The New York Times*, in which a group of nearly 20 prominent cognitive neuroscientists wrote:

> We know that it is not possible to definitively determine whether a person is anxious or feeling connected simply by looking at activity in a particular brain region. This is so because brain regions are typically engaged by many mental states, and thus a one-to-one mapping between a brain region and a mental state is not possible. (Lavazza and De Caro 2010)

This brings us to another crucial problem of localization: if brain qualia such as pain are to be reduced and decomposed to neural process, they must first be given a functional definition or else the reduction enterprise fails (Harley 2004; Kim 2006). Such a definition can only stem from a robust theoretical basis. However, there is no theoretical (psychological or philosophical) foundation for localization claims (Uttal 2002; Gerlach 2007; Poldrack 2010; de Graaf, Hsieh, and Sack 2012; Rathkopf 2013). The ability to infer about neural correlates of a cognitive process is often confined by the conceptualization of the process into a task that can be performed in the scanner (Bell and Racine 2009). Neuroimagers, while refraining from addressing the necessary psychological theoretical discussion, still, through their experimental design, reflect their own interpretation of specific cognitive functions (Burock 2009; Huber 2009; Roskies 2010). Many worry that such interpretations of functional decompositions demonstrate a naïve understanding of the cognitive processes underlying the performance of complex tasks (van Orden and Paap 1997; de Graaf, Hsieh, and Sack 2012; Aru et al. 2012). Some go as far as claiming that neurological data is irrelevant to cognitive psychology until a complete psychological theory has been established, at which point the neuroscientific data would be redundant (Harley 2004; Coltheart 2006, 2004; Loosemore and Harley 2010).

It is important to remember that the malleability of cognitive neuroscience theories is a direct outcome of the pliable nature of several psychological and psychiatric theories. Both disciplines have allowed some research within them to lax its scientific rigor to a degree where some warn of an intellectual crisis (Fava 2006)[6]. If the psychological theories guiding the localization project are themselves to be doubted, what stock can be afforded to the theories cognitive neuroscience develops using functional NIs? (John,

6. Psychology's proclivity for severally exaggerated high rate of positive results is well documented (Fanelli 2010; Francis, Tanzman, and Matthews 2014).

Loewenstein, and Prelec 2012; Yong 2012). By using a concrete example (Kanwisher et al.'s localization of facial recognition to the brain region known as FFA), Mole et al. (Mole et al. 2007) claim that such studies are simply not new or illuminating: "It has told us that if there are special resources for the processing of faces then the FFA is a likely site for them. But scanning can do nothing to answer the question of whether there are such resources".

Localization is a particular case of reductionism, and therefore must be contextualized as such[7]. In the attempt to analyze complex concepts using NIs, a dramatic simplification and reduction of the study objectives must occur in order to attain a feasible experimental protocol (Huber and Huber 2009). This deflation is interesting not solely in terms of the scientific process, but also because it may pertain directly to a possible human cognitive feature that prefers simpler accounts, which could explain why reductionism appeals to us more strongly than holistic or complex accounts (Rose 1999; Bunzl, Hanson, and Poldrack 2010).

In lieu of the above, it would seem that localization is more of an experimental choice than a scientific model, a way of making sense of this entangled mass called the brain. There appears to be a tacit agreement within some circles of the cognitive neuroimaging community that the claims hatched within it are best described as heuristic placeholders. Such voices acknowledge that while localization is limiting, it is still an important step that, once corrected, can lead to better understanding (Mundale 2002; Bechtel 2002, 2004; Craver 2005). If localization is merely such a tool (one with experimental advantages but not necessarily a reflection of actual brain activity), then it brings to the surface the tension between knowledge shared within the community and that disseminated outside of it. This begs the ethical question that if a discipline knows that it employs tools for heuristic purposes, then why does it insist on knowingly generating public claims too strong to be supported?

Standardization

Another problematic methodological tool employed in neuroimaging is a consequence of the biological reality that no two brains are alike, neither anatomically or functionally (Miller et al. 2012). Despite more than a century of research, there is still no consensus on reliable delineation of functional subdivisions in the brain, mostly

7. Reductionism has generated a by now insipid platitude of interpretations far beyond the scope of this paper. Readers can enrich their knowledge of reductionism by referring to the droves of publications discussing it.

because there are no binary abrupt transitions between brain regions (Peterson 2003; Haueis 2012; Cox et al. 2014)[8]. This has brought about the practice of standardization: the process of either averaging out the scans of a given study or comparing all subjects from the same experiment to an established standard atlas. Both forms of standardization are precarious since an averaged brain represents all brains and none, akin to Quetelet's *l'homme moyen*. In addition, researchers rarely provide information as to the way the norm was created (i.e. the characteristics of the subjects used to that end; (Reeves et al. 2003)). Furthermore, some of the most popular atlases used for standardization are not only many decades old but also based on the anatomy of a single subject (Bogen 2002). Thus, the epistemic value of NIs depends upon whether the idealized brain it portrays is representative of real brains with regard to the anatomical, physiological, and psychological factors relevant to the question under investigation (Bogen 2002; Uttal 2013).

Going beyond its legitimacy, standardization distorts structure/function analyses: in the processing phase of creating a NI, brain regions with different functional profiles near the region of interest are averaged together across individuals, reducing both the resolution and the sensitivity of subsequent functional analyses (Saxe, Brett, and Kanwisher 2010). Thus, function cannot be assigned purely on the basis of spatial patterns (Sadaghiani et al. 2010). Unfortunately, this is too often left out of neuroimaging discussions (Jbabdi, Sotiropoulos, and Behrens 2013), and is utterly absent in mass media reports.

Processing

For cognitive neuroscience to generate general observations concerning cognition and behavior it must compute correlations across subjects (Roskies 2008), and then each individual brain scanned has to be mapped onto an average brain. For this, the raw time series must undergo preprocessing to reduce noise. Vul & Kanwisher (Vul and Kanwisher 2010) describe the highly complex process required to convert raw data into publishable NIs:

> The time series of voxel changes may be motion-corrected, coregistered, transformed to match a prototypical brain, resampled, detrended, normalized, smoothed, trimmed (temporally or spatially), or any

8. Once we throw in pluripotency, we critically restrict our ability to distinguish what regions do as a whole and what sub-regions do individually (Bogen 2002).

> subset of these, with only a few constraints on the order in which these are done. Furthermore, each of these steps can be done in a number of ways, each with many free parameters that experimenters set, often arbitrarily. After preprocessing, the main analysis begins. In a standard analysis sequence, experimenters define temporal regressors based on one or more aspects of the experiment sequence, choose a hemodynamic response function, and compute the regression parameters that connect the BOLD signal to these regressors in each voxel. This is a whole-brain analysis, and it is usually subjected to one of a number of methods to correct for multiple comparisons… the whole-brain analysis is often the first step in defining a region of interest in which the analyses may include exploration of time courses, voxelwise correlations, classification using support vector machines or other machine learning methods, across-subject correlations, and so on. Any one of these analyses requires making crucial decisions that determine the soundness of the conclusions.

This detailed description shows that BOLD-fMRI NIs represent mathematical constructs rather than physiological reality (Burock 2009). The abundance of mathematical processing applied to the raw data leads to a skewed representation and estimation of many neural activities directly pertaining to the cognitive processing of a given task. Thus, the nature of processing alone demands great caution in interpreting functional NIs in cognitive neuroscience context (Logothetis et al. 2001). Consider spatial smoothing as an example: after smoothing, each voxel contains a mix of its own signal and the weighted signal of surrounding voxels. The justification for averaging the BOLD signal over space is improving statistical sensitivity. At the same time, spatial smoothing generates a systematic bias of spatial localization (Sacchet and Knutson 2013), as separate and distinct activations progressively blend into one another (Geissler et al. 2005). Stelzer et al. (Stelzer et al. 2014) argue that more than 90% of the post-smoothing signal at any given location originates from neighboring voxels, thus increasing the numbers of false positive voxels. These authors went as far as stating that due to spatial smoothing it is impossible link fMRI data with data from other neuroscience disciplines.

How Statistical Tools are Used

The convoluted process of generating a functional NI does not stop with mathematically transforming raw data via processing. Since raw data are an astronomical amount of numerical values in long time-series, it is imperative to perform statistical operations in order to convert them into images. However, statistics, paraphrasing D'Israeli, are the most mendacious of all lies, and since the choice of statistical tools has a direct and paramount impact on the resultant image and the conclusions that can be drawn from it. And the list of available statistical tools in cognitive neuroscience is impressively long (Carp 2012a, 2012b).

Before we delve into the role statistics play in BOLD-fMRI, we have to present two basic definitions. A type I error falsely rejects a true null hypothesis and generates a false positive: accepting that a hypothesized event exists when it does not (e.g. a wrong medical diagnosis). A type II error is the acceptance of a false null hypothesis, yielding a false negative. Importantly, false negatives are correctable with ensuing research, whereas false positives are difficult to refute once established in the literature and not re-evaluated properly (Bennett, Wolford, and Miller 2009). The various factors contributing to elevated false positive rates in BOLD-fMRI (Bennett, Wolford, and Miller 2009) can be brought under the umbrella of poor application yielding low statistical power, which in turn complicates replication and fosters contradicting conclusions in the analyses of the same database (Duncan 2001; Button et al. 2013; David et al. 2013).

A ubiquitous statistical error in functional neuroimaging is the non-independence error (aka double dipping): using the same data for selecting the voxels of interest and then using these voxels for the secondary analysis, the one upon which the functional conclusions are based[9]. Double dipping violates random sampling because the test statistics are not inherently independent of the selection criteria of the region of interest, thus statistically guaranteeing the outcome of the second analysis and rendering them useless (Kriegeskorte et al. 2009; Vul et al. 2009). Similarly, as mentioned before, statistical tests in neighboring voxels are not independent of one another, because time series in neighboring voxels are intercorrelated (Peterson 2003). Analyses have shown that the non-independence error is widespread in BOLD-fMRI studies (40-50% of published papers) and that the severity of the distortions of the results presented in these papers could not be assessed. This necessitates replications and reanalysis (Kriegeskorte et al. 2009) or the results of these studies "mean almost nothing", since they are "using

9. It must be noted that this problem is not unique to cognitive neuroscience, as it is widespread in neuroscience and psychology too (Fiedler 2011).

seriously defective research methods and producing a profusion of numbers that should not be believed" (Vul et al. 2009).

In addition to double dipping, there are other ill-used statistical tools in cognitive neuroscience. A major such source of complication is the necessity to correct for multiple comparisons: many researchers find these corrections too draconian, and choose to either avoid correction altogether, or to employ lenient statistical tools (Saxe, Brett, and Kanwisher 2010). Bennett et al. (Bennett et al. 2011) used an extreme test case to demonstrate that by using uncorrected statistics for multiple comparisons they were able to generate a NI that showed active voxel clusters in the brain of a dead fish in response to visual stimuli. The authors concluded that it is likely that "investigators do not want to jeopardize their results through a reduction in statistical power".

Another case in point is the statistical dichotomy between significant and non-significant results based on P values. This comparison often erroneously involves two separate tests in which researchers conclude that effects differ when one effect is significant ($P<0.05$) but the other is not ($P>0.05$), while the comparison should be between them (Nieuwenhuis, Forstmann, and Wagenmakers 2011). Numerous (and early) articles have clearly demonstrated that this dichotomy is arbitrary and unwarranted, as any strong evidence against a null hypothesis (if such at all exists) depends on other conditions and cannot be expected to be globally valid at $p<0.05$ (Sterne, Cox, and Smith 2001; Wacholder et al. 2004; Ioannidis 2005b). Nonetheless, this did not prevent the spread of the statistical error common in neuroimaging studies of comparing significance levels (Henson 2005; Poldrack et al. 2008). Nieuwenhuis et al. (Nieuwenhuis, Forstmann, and Wagenmakers 2011) found that this dichotomy is prevalent even in high profile journals, and that in some cases the error may have contributed substantially to the article's main conclusions.

Yet another sizeable statistical concern is unfitting sample sizes: most published fMRI studies have sample sizes that would be considered exceedingly small by conventional standards (Yarkoni 2009; Button et al. 2013; Ingre 2013), if they include sample size calculations at all (Guo et al. 2014). It is established that in fMRI studies, small studies (n=16) fail to reliably distinguish small and medium-large effect sizes from random noise as do larger studies (n=100) (Ingre 2013)[10]. However, Wager et al. (Wager et al. 2009) report that across 415 fMRI studies reviewed, the average group size was smaller than 12, with some using only 4 subjects. At the same time, the number of activation loci claimed to be discovered by them is relatively large (David et al. 2013).

10. Zandbelt et al. (Zandbelt et al. 2008) provide a sample size estimations for BOLD-fMRI crossover studies.

This statistical bungle is exacerbated by the observations that most labs employ statistical tools according to historical precedent rather than through formal power calculation (Button et al. 2013). These statistical methods were developed to allow fMRI to *detect* activation rather than *characterize* it, thus making the interpretation of results often speculative (Monti 2011). Unfortunately, most fMRI researchers have only a vague idea of how reliable their results are, and the more tasking cognitive attributes are the ones with the lowest fMRI reliability (Nichols and Hayasaka 2003; Bennett and Miller 2010; Saxe, Brett, and Kanwisher 2010). In fact, Uttal (Uttal 2013) argued that "many statisticians would be amused by the cavalier attitude of some neuroscientists in assuming that their data meet the most basic criteria for statistical robustness" (p. 55). Ioannidis (Ioannidis 2005b) lists six parameters diminishing the probability that statistical findings in functional NI are valid: 1) small studies; 2) small effect size; 3) the greater the number and the lesser the selection of tested relationships; 4) high design and analysis flexibility; 5) financial stakes and other biasing elements; 6) a hot field drawing many labs to it. Cognitive neuroscience falls short on all these criteria, thus casting a looming shadow over their produced claims.

Another crucial example of lenient scientific austerity, a direct amalgamated result of the methodological shortcomings listed above, is the alarmingly low rate of experimental replication in both psychology and fMRI studies (e.g. (Pashler and Harris 2012)). Moreover, when replication does take place it often contradicts initial reports, particularly if those were based on small samples size and/or published in high impact factor journals (Ioannidis 2005a). As shown earlier, conclusions drawn in studies committing methodological and statistical errors can continue to propagate and serve as basis for future null hypotheses because older studies are rarely re-evaluated and the publication process is biased toward positive results. The accumulative effect of many such distortions, regardless of their magnitude, is a grave impact on the validity and robustness of localization claims.

The wide range of tools and analyses that can be operated on the full arsenal of methods applied in BOLD-fMRI research leads to a widespread phenomenon throughout science, that of high analytic flexibility and selective analysis reporting: choosing the most favorable experimental/analytical combination, the one that promote positives results (Carp 2012a; Button et al. 2013).

Variability

The final methodological concern is results variability, which runs the gamut from within- to between-subjects and between labs. The obstacle is that such variability is yet another variable hampering the replication and attainment of consistent results (Uttal 2013). Early fMRI studies have documented intra-subject variability, even after repeated tests in the same laboratory over a number of days (Zandbelt et al. 2008), particularly with cognitive tasks (McGonigle et al. 2000). Later work has unveiled inter-subject variability, and showed it to be greater than the intra-subject one (Miller and Van Horn 2007; Miller et al. 2009; Diederen et al. 2013; Tancredi and Brodie 2007). Additional variability exists between different laboratories: while individual experiments identify only a relatively small number of activation peaks per cognitive task, collecting all responses across many centers tackling the same cognitive attribute generates a distribution map covering the entire brain. Inter-venue variability is apparently so great that meta-analyses only exacerbate the situation and increase variability (Uttal 2013; Fox et al. 2015). The cumulative effect of these types of data variability is a serious impediment on the localization project, suggesting that there are no macroscopic-level delineations corresponding to cognitive performance, and that they are probably a methodological artifact (Gonzalez-Castillo et al. 2012; Thyreau et al. 2012).

Summing up, this overview of central and ubiquitous BOLD-fMRI methodologies demonstrates that the experimental design of functional neuroimaging studies (in addition to the restrictions imposed by the technology itself) acutely delimitates strong localization claims for pinpointing the neural substrate of cognitive functions. While some technological and methodological advancements have presented themselves throughout the years, they had alleviated mostly minor concerns. Methodologically speaking, the experimental rationale has remained mostly intact, and the philosophical concerns - which constitute the very keel of cognitive neuroscience's arguments - still linger on and represent inherent flaws looming large over the validity of the localization project. Nonetheless, the allure of fMRI has attracted many scientists from different disciplines to use it in their work, and too many of them prematurely capitalize on established protocols rather than addressing their particular scientific needs (Pan et al. 2011). Thus, a growing number of neuroimagers are nescient with respect to the complexities and problems associated with BOLD-fMRI and the inner workings of MR machines and their capabilities. At the same time, the physicists and mathematicians responsible for improving MRI technology lack an intimate knowledge of cognitive and neurological

theories (Peterson 2003; Seixas and Ayres Basto 2008). This yields a dialogue of the deaf between producers and users, a situation not conducive to proper scientific practice.

BOLD-fMRI - Philosophical Issues

As stated in the opening of this paper, the problems associated with the BOLD-fMRI NIs for the purposes of cognitive neurosciences stem from a quadrumvirate of levels, each posing difficulties that are hard to alleviate. We have seen the kind of difficulties associated with the BOLD signal itself and the technological parameters of the MR machinery. Then we discussed experimental design and statistical manipulations and learned that they are significantly harder to allay. In this next tier, I focus on concerns emanating from philosophy of mind and philosophy of biology that constitute the most tenacious opposition to the localization project.

The Mereological Problem: Psychophysics Revisited

Neuroscientists are usually materialists that vehemently deny there is more to the mind than what the brain has to offer. The hubris of neuroscientists reflecting utter confidence in their ability to solve all things mind is best exemplified by Francis Crick's statement "No longer need one spend time attempting… to endure the tedium of philosophers perpetually disagreeing with each other. Consciousness is now largely a scientific problem" (Crick 1996).

When neuroscientists claim to have discovered the neural correlate of a cognitive trait, the fundamental question from philosophy is what is it really that they show. For such a claim to be adequate, an isomorphism between neuronal form and function and experiential content, at a not-established description level, must exist (Noë and Thompson 2004). As demonstrated earlier, such a one-to-one correspondence, at least via functional neuroimaging, is currently not even technologically feasible. Furthermore, philosophers seriously doubt the validity of such future argument, even if better technologies were to present themselves (Sprevak 2011). Their counter argument is that feelings are felt, experiences experienced, thoughts created, and behavior displayed only at the level of the whole person interacting with her environment (Noë and Thompson 2004; Burock 2009). It is the person that cognizes, not her sub-personal organs, tissues, cells, organells, or molecules; not even if they are called the brain, the cortex, the amygdala, neurons, synapses, glial cells or dopamine. This is the mereological fallacy: assigning function to a part of a whole that is attributable only to the whole itself (Bennett and Hacker 2003; Pardo and Patterson 2010).

The psychophysical and mereological problems boil down to the vernacular question "are we our brains?". As we have seen in the discussion of localization, reductionism has been the scientific bon ton for many decades now. Neuroscience offers some aggressive forms of reducing the mind to the brain, such as equating mental processes with neural processes or arguing that mental processes are causally inert epiphenomena of neural processes (Beauregard 2009). Fortunately, not all philosophers or neuroscientists subscribe to these points of view. Some philosophers reject reductive materialism by arguing that the brain cannot participate in the sensory and in the social (Burwood 2009), and that mental processes exert a causal influence on the brain (Paquette et al. 2003; Beauregard 2009). Recent scientific work strongly suggests that the body affects the development, homeostasis, and plasticity of the nervous system (e.g. (Qureshi and Mehler 2013)). Such findings have led some to argue that neuroscience is, ironically enough, a dualist enterprise: while rejecting a dichotomy between brain and mind, they ignore somatic effects on the brain, thereby effectively creating a dichotomy between the brain and the body (Glannon 2009).

If the rebuttal of a mandatory body-brain-mind-Umwelt complex is valid (Chiel and Beer 1997; Byrge, Sporns, and Smith 2014), then not a single component of this complex is sufficient by itself to generate and explain cognition and consciousness, as each carries only a proportional weight within that complex (Glannon 2009; Pardo and Patterson 2010; de Graaf, Hsieh, and Sack 2012). A useful analogy comes from another field that was dominated by fierce reductionism: genetics. Maybe if we think in terms of genotype (the brain, the neurotype?) and phenotype (the mind, the cognitype?) we could better explain why neuronal operations are necessary but not sufficient to explain the mind. As such, there is no 1:1 correspondence between neurotype and cognitype, as the neurotype serves only as a scaffold upon which the cognitype builds and elaborates via dynamic reciprocal interactions with the body and the environment. If we combine this with the persistent criticism expressed against cognitive neuroscience's lack of psychological and cognitive theoretical background ((Uttal 2002); see page 6) we end up with Coltheart's (Coltheart 2004) statement: "No amount of knowledge about the hardware of a computer will tell you anything serious about the nature of the software that the computer runs. In the same way, no facts about the activity of the brain could be used to confirm or refute some information-processing model of cognition" (p. 22).

Form and Function

The study of the dynamic relationship between form and function and the constrains they levy on each other has fascinated biology for centuries (Mundale 2002; Wouters 2005). Neuroscience is no exception and contributes its own questions to this debate (e.g. (Meinertzhagen et al. 2009; Friston et al. 2010)). Cognitive neuroscience presents a philosophical challenge because it demands that we explicitly define the form and function of consciousness in order to be able to properly design an experiment that will accurately identify its neural loci. Such definitions break down to questions such as what is a brain function in general or what is the function of a given brain area. Lamentably, this brings us back to the severe lack of theory in functional neuroimaging research. The form-function relationship is probably the crux of the difficulty functional neuroimaging has in establishing its claims: as long as there is no full mapping of what functions are served by which brain region and as long as the boundaries of these regions cannot be precisely delineated, no trustworthy localization claims can be made. Things get murkier when we remember that NIs address multiple neuronal levels of organization, each characterized by different expressions of form and function. Therefore, for functional neuroimaging to have a legitimate seat in the mind/brain debate, it has to meet with two prerequisites. First, at the very minimum, there has to be a clear definition of what brain/cognitive functions are for each and every experiment. Second, an explicitly detailed account of the neural substrate, be it localized or distributed, of that function must be given. However, this harks back to the problem that by answering these questions, any additional neurocognitive data would be redundant (see page 5 and 11).

After touching upon some philosophical concerns related to the neuroimaging practice of localization and its use in establishing claims about the nature of the brain-mind link, glaring deficiencies in the theoretical foundation of this scientific pursuit appear. Unlike the technological and methodological aspects discussed earlier, the philosophical lacunae are very difficult to alleviate, especially without overhauling the entire discipline. Therefore, it seems unlikely that a major revision of the core tenets of NIs' use in the study of cognition will present itself in the near future. A corollary of this observation is that the scientifically dubious knowledge gained so far from this discipline will continue to proliferate unbridled from the corridors of academe through to office buildings coolers.

As an interim conclusion, I describe functional Nis' inadequacy to support localization by alluding to the degrees of separation a NI has from the biological phenomenon it allegedly represents. The **first** degree of separation stems from the nature of the BOLD signal itself: the tremulous nature of the neurovascular coupling hypothesis prevents a

definite access to the neural activity measured. A **second** degree is introduced owing to the technological parameters of the MR scan, which force a spatiotemporal dissonance with the neural activity measured. Methodological choices force the next degrees of separation: the **third** separation results from violation of the pure insertion assumption, which then prevents us from linking a specific cognitive attribute to a brain region. A **fourth** degree enters the equation via methodological shortcomings, ranging from the failure of reverse inference, through the lavish baggage of processing, statistical manipulations, standardization and culminating with all the processes designed to generate a smoother and cleaner image by discarding or hiding data that does not fit well with the a posteriori assignment of the region-cognitive function link promoted in a given study. These degree force us to sincerely doubt the validity of the localization endeavor. Finally, the **fifth** degree results from the aesthetic predilection of neuroimagers (see p. 13). Thus, owing to these degrees of separation, NIs presented as a depiction of cognition and consciousness actually have only a gossamer tenure with neuronal reality.

Ethical Considerations

The ultimate goal of this paper is not merely to map the spectrum of problems associated with the use of BOLD-fMRI NIs use in cognitive neuroscience. Rather, it is to use this map in support of the argument that this practice all too often amounts to unethical science, one where the generators of data overlook known shortcomings of their tools of the trade and press forward with producing claims too strong to be supported by exploiting the strong appeal of their meticulously crafted images. These claims filter through and find their way to non-professionals and policy makers who are oblivious to the tangled web of complexities surrounding these captivating images and who lack the tools to doubt the conclusions attached to them.

Having discussed the scientific frailty of using NIs for studying consciousness and behavior, in this section I focus on two aspects of neuroimaging practice bearing ethical impact. First, I establish and explore the implications of the absence of a theoretical framework linking neuroanatomy with all things mind and self. Second, I investigate the role representation plays in the perception of functional NIs and the propagation of the messages they attempt to convey.

Lack of Theory

Ample examples were given throughout this paper (e.g. localization, p. 6) and are given in this section as to the understanding, covering both proponents and opponents

of neuroimaging's use in cognitive neuroscience, that the field lacks a theoretical background to lean on. This refers to the absence of both a consensual psychological framework of cognition (van Orden and Paap 1997; Uttal 2002; Harley 2004; Coltheart 2006; Poldrack 2010; de Graaf, Hsieh, and Sack 2012; Klein 2012; Rathkopf 2013; Reiner 2011; Fox and Friston 2012) as well as a well-established unified theory of how the brain works (e.g. (Haxby 2010; Power et al. 2010)). Let us begin with a fundamental conundrum, a critique of physicalism (the identification of mental states with brain states): if a creature without a brain can think, thinking cannot be a brain state (Block 1996). To answer this intellectual exercise we must define what a brain is and what constitutes a thought. There are vast and profound differences between the mammalian, piscine, and insect brain, and still some insect and avian species outperform some mammal species in various cognitive tasks (e.g. (Shifferman 2011)). So, what is a brain? Is it the highly distributed and restricted nervous system of the ant or bee? Would the nervous system of cephalopods qualify? What is the neuronal communality that allows profoundly different nervous systems to still generate the (seemingly) same behavior, and how could it be such simple such systems outperform advanced systems in particular tasks?

This has led Uttal (Uttal 2002) to argue that it is impossible to define the cognitive attributes to be localized without circularity and imprecision, which, in turn, inevitably lead to erroneous localizationist claims. It also brought V.S. Ramachandran, a prominent psychologist, to opine that "98% of brain imaging is just blindly groping in the dark" (Dingfelder 2008). The lack of theory is further aggravated by the fact that we are constantly learning new things about very fundamental aspects of brain function[11]. A prime example of the combination of both the lack of theory and the constantly developing body of knowledge is the case of the body-brain link discussed earlier (p. 10).

Another bias in neuroscientific research that weighs heavily on the validity of localization conclusions is the focus on event-related activity, knowingly overlooking spontaneous neuronal activity, the intrinsically generated brain activity that is not attributable to specific stimulus (Fox and Raichle 2007). The average adult human brain consumes 20% of all the energy consumed by the body, yet event-related induced energy consumption accounts to less than 5% of the baseline level of activity (Raichle and Mintun 2006). Thus, to understand the brain we must overcome this metabolic bias and consider the component that consumes most of the brain's energy: spontaneous neuronal

11. A very recent random yet significant example is that different neurons have different profiles of longitudinal myelin distribution, thus directly shaping its communication range and abilities (Tomassy et al. 2014).

activity (Sadaghiani et al. 2010; Fox and Raichle 2007). Spontaneous brain activity fluctuates within and between different modes and should not be considered noise, as it is coherently expressed in larger neuronal populations and functionally meaningful (Laufs et al. 2003). Perhaps the most studied example is the default mode network, an anatomical assembly of multiple brain regions supporting the "stand by" state of alertness (Raichle et al. 2001). It is argued that this base level is designed to maintain and support a dynamic shift between an introspective, self-referential mode of mental activity to an extrospective, preparedness mode that remains alert to environmental changes (Fransson 2005). Thus, rest is a state of a continuous orchestrated activity that is intermittently overridden once a goal-oriented activity emerges in response to certain stimuli (Fransson 2005), which persists through active cognitive task performance (Fox and Raichle 2007). Hence, before we bridge this knowledge gap and understand better spontaneous neuronal activity, no localization argument of substance can be made.

Augmenting the deep problem of lack of theory in the design and analysis of functional NIs is the fact that other disciplines within neuroscience present alternative and opposing interpretations of the neural basis of cognition. One can imagine here a pincer movement, wherein the cognitive neuroscience narrative of cognitive functions sequestered to specific brain regions is challenged simultaneously but differently from both top and bottom. From the lower organization level perspective, some shift the reductive fulcrum to single cells (individual neurons or groupings of identical neurons) and argue that they alone suffice to support the execution of some cognitive functions [e.g. (Smith and Ratcliff 2004; Nieder and Merten 2007)][12]. The attack from north points to the brain as a highly interconnected, spatiotemporal dynamic system that uses distributed representational schemes and relies on contextual and often transient sharing of neural resources across tasks in asynchronous and parallel fashion (Fingelkurts, Fingelkurts, and Kähkönen 2005; Fox et al. 2005; Henson 2005; Nair 2005; Sporns 2014; Zeki 2015). Proponents of networks as the foundation of cognition argue that there is extremely limited evidence supporting local non-linear neuronal operations, and these examples can be explained alternatively by using higher-level neuronal elements and emergence (e.g. (Bermúdez i Badia, Bernardet, and Verschure 2010)). They further assert

12. A much studied and criticized example of single cell-based cognition is grandmother cell theory. These are neurons argued to be solely responsible for the neuronal response for a stimulus, as they respond only to a very specific stimulus using neural convergence, in which neurons compute their various inputs in to a complex representation of a specific percept (see (Gross 2002; Quian Quiroga et al. 2008; Quian Quiroga et al. 2005)).

that the composite downstream effect of these neuronal assemblies (i.e. cognition), cannot be achieved either by single neurons alone (Buzsáki 2010) or by a brain region (Petersen and Fiez 1993).

Clearly, these accounts are mutually exclusive, thus reiterating the dire need for a comprehensive neural theory consolidating the abundant neuroscientific and psychological data and models. Also evident from glimpsing other disciplines of neuroscience is that a single neuronal element cannot, by itself, be a sufficient explanation either for the operation of neuronal systems or for the cognition/behavior it supposedly supports. If that were the case then the function of the entire nervous system would have collapsed into the operation of that single element, thus nullifying the need for a system. In that neuroecological context, functional NIs fail as they ignore both lower and higher organization levels: at the nerve cells level it is technically blind to a rich arsenal of microcircuitry, while at the regional level it overlooks the fact that there is no simple correspondence between structural and functional domain boundaries (Damoiseaux and Greicius 2009). This state of affairs has led Reiner (Reiner 2011) to assert that

> Until we have a satisfying mechanistic account of how two similar neurons become distinguishably specialized in their selectivity it is best if arguments concerning how brains work adopt a dash of reservedness, one that realizes the limitations of our technology and method and acknowledges that all we have is a horde of observations in search of contextualization and deeper explanation.

Representation: How neuroimages are perceived and interpreted outside the lab

> Nowadays, the greatest conceptual abstraction is to be found in conceptual images… the greatest imagination is to be found in scientific texts. Thus, behind one's back, the hierarchy of codes is overturned. Texts, originally a metacode of images, can themselves have images as a metacode (Flusser 2000).

I turn now away from the producers of NIs and investigate how the claimed (by neuroimagers) and marketed (by press officers and mass media) visually supported arguments about the mind-brain connection compare with what is actually perceived outside the lab. Several studies have highlighted the significant weight visual evidence carries over non-visual evidence, such that an image has a greater heuristically persuasive

power and is deemed to be an even more accurate representation of a given phenomenon than are statistical and numerical presentations (Dumit 2004).

The fact that an image embodies some information does not suffice to account for its representational content. Since both referents and contents can be assigned by stipulation, just about any object can be used to stand for anything (Goodman 1976; Roskies 2008). Reiterating the technological and methodological lacunae presented earlier, Perini (Perini 2012) has pondered how NIs support scientific claims if their generation is not a simple matter of nature re-presenting itself legibly. Perini argues that NIs are not mimetic in the way photographs are since what they allegedly represent (location and level of neural activity) are not visual properties. This means that comprehending an image as a representation of something else always involves a kind of interpretation, and this representation hinges on shared interpretive practices (Perini 2012). NIs are presented not only as (at best) a substitution (i.e. a hypothetical construct) or as an epistemological or heuristic scientific observation, but rather as an actual phenomenological realization of the brain (Huber 2009).

Clearly, not only do non-professionals and neuroimagers not share the same epistemological field (which would facilitate effective communication between them), they are also separated in their epistemological status and roles (non-professional rarely can contest scientists' claims without the assistance of other scientists). Roskies (Roskies 2010) concluded that functional NIs' epistemic status rests on inferential distance: the actual biological phenomenon studied is inferentially far removed from the images themselves. Thus, a NI has a strong impact on the viewer while having limited scientific content. Roskies sets apart actual inferential distance (inferences inherent to the scientific procedure) from apparent inferential distance (the confidence non-professionals have in the scientific conclusion based on the NI). When these inferential distances come apart, people are prone to assign an unwarranted epistemic status to scientific claims. This means that NIs take on evidential roles not as a direct representation of natural phenomena, but only as the result of activities aimed at assigning referents and attributing content to them. These activities, in turn, are not scientifically objective and are heavily influenced by cultural and social norms, and we have to decipher and unfurl these norms if we are to see beyond the mediation of the visual (Lynch 1991; Joyce 2008; Burri 2012).

An example of a tacit scientific culture that fuels inferential distance is that NIs have by now come to constitute an aesthetic (Burri 2013; Aguirre 2014), and an affordable one at that since achieving a "standardized" aesthetic is made easy with the availability of free statistical parametric mapping software that perform the above processes and

allows for push-button analysis of neuroimaging data with minimal understanding of the many statistical processes and assumptions (Aguirre 2014; Joyce and Hayasaka 2012; Lynch 1991). To highlight this point, Burri (Burri 2012) quotes a sales manager of an MRI scanner manufacturer who described neuroimaging conferences as follows: "It is like a beauty contest... You must see beautiful images that are high in resolution, that are luminous and perfect".

In that regard, Frow (Frow 2012) investigated digital image processing guidelines of leading contemporary interdisciplinary science journals. These guidelines were put in place in order to detect inappropriate image manipulation *after* an image had been captured. Most guidelines focus on two concerns which pertain directly to the critique of functional NIs. The first requires that any adjustment made using digital processing must be applied to the whole image rather than selectively to specific parts of it. Secondly, adjustments that obscure or remove information from the original image are forbidden. Frow asserts that these guidelines reflect both a desire to redefine acceptable and unacceptable practices in image production, as well as a pressure to produce ever more visually appealing images to embellish journal covers. This approach might be interpreted as hypocritical given the statement of *Nature*'s editor in these guidelines that "beautification is a form of misrepresentation. Slightly dirty images reflect the real world"[13].

Another interesting point raised by Frow is that these guidelines also serve the goal of protecting the scientist as a skilled professional: before the advent of digital image processing, the creation of scientific images required a substantial level of technical mastery, but the need for such expertise has long evaporated thanks to photo editing software. Thus, Frow's study highlights a double standard: while image processing of certain data is now unacceptable in some esteemed circles due to fear of excessive manipulation, a significantly more excessive (and often not accounted for) manipulation is celebrated in other circles.

The aesthetic angle reveals a tension between what is known scientifically and what is presented publicly, but how much of this discussion is relevant to the general public? Which scientific news make it to popular media? Suleski & Ibaraki (Suleski and

13. A swift, back of a napkin inspection of the author guidelines of the six highest impact factor journals in the category of radiology, nuclear medicine, and medical imaging according to Thomson's Journal citation reports (Human Brain Mapping, Radiology, NeuroImage, JACC: Cardiovascular Imaging, Circulation: Cardiovascular Imaging, and Journal of Nuclear Medicine) has shown that these journals do not address image processing in their guidelines as of the date of submission of this MS.

Ibaraki 2010) show that roughly only a permille of published scientific papers reach the mass media, with health/medicine papers taking the lion's share of coverage. This miniscule sliver from the humongous scientific output is a direct result of a phenomenon dubbed scientific sensationalism: a complex process in which very specific bits of scientific knowledge are aggressively pushed to the forefront of mass media to receive a disproportionate piece of the public's attention. Sensationalism has several ingredients that are important for understanding the success of functional NIs. First, a conflict of interests and a bilateral miscommunication between academia and mass media as well as differences in reporting style typical of each (Ransohoff and Ransohoff 2001; Woloshin and Schwartz 2002; Rose and Abi-Rached 2013)[14]. Second is a bias shared by both academe and mass media for publishing predominantly positive results while omitting negative ones, thus skewing both scientific and public perception (Easterbrook et al. 1991; Koren and Klein 1991; Cassels et al. 2003; Zuckerman 2003; Caulfield 2005; Brechman, Lee, and Cappella 2009; Gonon, Bezard, and Boraud 2011)[15]. This bias is well documented in functional NI as well (Ioannidis 2005b; John, Loewenstein, and Prelec 2012; Vul and Pashler 2012; Ioannidis et al. 2014) Third, mass media tend to flatten scientific reports and strip them of the many complexities that characterize them (Woloshin and Schwartz 2002; Beck 2010; Schwartz et al. 2012). Combined, these phenomena generate hype fluctuations that misinform the public and cause it to doubt scientific results (Ransohoff and Ransohoff 2001), and produce errant cultural residues (Conrad 1997; Gonon, Bezard, and Boraud 2011).

Evidently, neuroscience and neuroimaging, being part of the health world are no exception and experience sensationalism (O'Connell et al. 2011) and aggressive commercialization too (Chancellor and Chatterjee 2011). Barring highly irregular cases, the flow of information from lab to media is characterized by the ironing out of technological and methodological concerns, the discard of cautionary comments and the omission of alternative explanations, all resulting in distorted conclusions committed to popular memory. If these were merely misconceptions rampant within the community

14. As an example, Robillard & Illes (Robillard and Illes 2011) report that nearly half of the neuroscientists they have interviewed claimed that their academic institutions frown upon their efforts to communicate their research to the public.

15. Sensationalism can escalate the commercialization of academia (Downie and Herder 2007; Hong and Walsh 2009) and even bias scientific practice: fMRI studies were cited three times more often than lesion studies of the same brain region, mostly due to the fact that they were published in higher impact factor journals (Fellows et al. 2005).

of neuroimagers it would amount to scant scientific practice. The problem is different: while they *may* know better, what they *communicate* to the outside is not. That equals unethical practice.

An example of scientific sensationalism at the academic level can be seen in Charest et al. (Charest et al. 2014) fMRI study of semantic space. In the Significance section of the article, the authors write "our results demonstrate that fMRI has the power to reveal individually unique representations of particular objects in the human brain. The novel method might help us understand the biological substrate of individual experience". However, in the concluding paragraphs of their discussion they write "It is important to note that the predictions of perceptual idiosyncrasies from the hIT representation, although robustly better than chance, are not very precise. Precision estimates depend on many factors, and have little meaning beyond the context of a particular study".

An example of scientific sensationalism in mass media is a famous article published in *The New York Times* on 11/11/2007 titled "This is your brain on politics". In anticipation of the 2008 US presidential elections, the article described an fMRI experiment in which twenty swing voters were scanned while images and videos of candidates were presented to them. The authors concluded that the brain responds differently to Republican and Democrat candidates as well as to the words "republican" and "democrat" themselves. They even ventured that voters had mixed feelings toward Hilary Clinton, while Mitt Romney showed potential. The article generated instantaneous political and cultural fervor, and that of the scientific community soon followed as it had realized that this attempt to impact the results of the elections was not only egregious in intent, but also flagrantly scientifically vacuous. One of the leading cognitive neuroscientists, Russell Poldrack commented:

> It was really closer to astrology than it was to real science… it epitomized everything that a lot of us feel is wrong about where certain parts of the field are going, which is: throw someone in a scanner and tell a story about it… people will start to see fMRI as neophrenology, just telling stories and not giving explanations (Ramani 2009).

The elections story becomes even more nefarious when considering a meeting in 2005 that brought together leading neuroscientists, ethicists, and journalists to discuss various aspects of neuroimaging[16]. Participants were of the opinion that neuroscientists have a

16. *Hard science, hard choices: Ethical questions & public policies for the emergent science of the brain*. 10-11/05/2005, Library of Congress and The Dana Foundation, Washington, DC.

responsibility to clarify to the public what are the limitations of cognitive neuroscience research. Some even argued that scientists are obliged to be more rigorous in their research (Check 2005).

The elections study has led researchers to investigate directly the effect NIs have on layperson. This work has yielded conflicting results: some argued that part of the credibility afforded to brain imaging lies in the image itself (McCabe and Castel 2008; Roskies 2008; Keehner and Fischer 2011; Ikeda et al. 2013; Saks et al. 2014), while later work reported no such special effect (Gruber and Dickerson 2012; Hook and Farah 2013; Michael et al. 2013; Schweitzer, Baker, and Risko 2013). Interviews conducted with neuroimagers, patients, and other non-professionals show that NIs are performative and enact the body rather than transparently represent it (Joyce 2005; Casini 2011). The performative can also easily cross the line and become deceptive, as evidenced by quotes from MRI technicians who noted that "It's easy to tweak the parameters to make something that's not there" or admitted that MRI images are all "smoke and mirrors" (Joyce 2005). Burri (Burri 2013) unravels similar observations from her interviews: when initially asked to describe MRI images, scientists portray them as a "document which depicts reality 1:1"; "it's a photograph"; or "if you would slice the body, it would look exactly like that". The flip side is given via a psychiatric patient who had undergone fMRI and stated that "It's a picture of who you really are. On the inside" (Cohn 2010). Burri (Burri 2013) quotes a patient stating that "the image is... something irrevocable", echoed by a physician who said that "The images persist. Patients remember them well". Another neurophysiologist recounted presenting at a conference and admitted "We presented [the images] in a really wrong way. [The audience], however, liked it. People were not aware that the images were wrong". However, when asked for a minimal level of reflection, the quotes from Burri's interviews quickly change to "there is a danger in the images. Because images sometimes suggest more than they should", or "in every image there is something delusive". Alarmingly, that is not the full extent of it. Burri quotes a radiology professor claiming that "images pretend a lot of authority, seeming authority that absolutely doesn't exist". This state of affairs has led Carp (Carp 2012a) to claim that "A motivated researcher determined to find significant activation in practically any brain region will very likely succeed – as will another researcher determined to find null results in the same region".

Expressing his opinion of the state of neuroscience, Martell (Martell 2009) was very adamant in stating that "the ability of neuroscientists to use neuroimaging reliably to predict (and perhaps… postdict) thoughts or behavior is currently nil". Wolpe (Wolpe 2006) stated that "science has become one of the most powerful and pervasive forces

for change in modern societies. As the professionals at its helm, scientists have a unique responsibility to shepherd that change with careful ethical scrutiny of their own behaviour and thoughtful advocacy of scientific research". Similarly, Lavazza & De Caro (Lavazza and De Caro 2010) assert that "When one comes to the issue of human agency, great caution should be used before drawing bold philosophical, political, and social conclusions from neurological findings, whose correct interpretation and value are still extremely controversial". To add insult to injury, research shows two worrisome trends in public's perception of the scientific explanation of mental illness. First, the discourse is dominated by biological models. Second, and as an upshot, reductionist causal explanations tend to exacerbate negative feelings toward the mentally ill (Angermeyer et al. 2011; Schomerus et al. 2012; Kvaale, Gottdiener, and Haslam 2013; Lebowitz and Ahn 2014). This exact public reaction has been documented previously in the case of genetic explanations in general and those of brain/mind in particular (Dar-Nimrod and Heine 2011; Haslam 2011). Rectifying the epistemological limitations of fMRI is an ethical imperative on two levels: firstly, because they constitute a threat to the quality of research (Kaposy 2008; Anderson, Mizgalewicz, and Illes 2012; Bluhm 2013; Peterson 2003); and secondly (and perhaps more importantly) because they perpetuate a skewed misconception of what brains are and do as well as what their connection is to cognition, mind, and the self.

Nonetheless, it must be remembered that the public, though not well versed in the particulars of neuroimaging, is far from inert: people exercise critical judgment of scientific news and incorporate various discourses to form an opinion (Wynne 2001; Meurk et al. 2014). Furthermore, the public must not be seen as a monolithic entity, but rather as enjoying different perceptions of science, brain, and mind (Cohn, Dumit, and Roepstorff 2003). However, a lingering question is have audiences already been numbed by NIs in popular media? Whitely (Whitelely 2012) examined mass media publications between 2005 and 2009 and learned that coverage of fMRI studies has often substituted NIs with artistic renditions. The question is whether NIs have thus become synonymous to brain and cognition or have they become void of meaning beyond their aesthetic value and reduced to a typological brand name. This visual saturation scenario might explain why studies published from 2013 onward failed to support the argument the NIs impact viewers in a unique fashion. Another explanation comes from Fernandez-Duque et al. (Fernandez-Duque et al. 2014), who show that neuroscientific explanations (be they accompanied by NIs or not) carry a distinct allure bestowed with exaggerated credence which amounts to a conceptual bias. Vidal (Vidal 2009) provides a detailed anthropological account of this phenomenon and dubs it the brainhood, or the cerebral subject ideology.

In conclusion, it is safe to say that within the scientific community, neuroimagers reject NIs as a visual technology and categorize it instead as a numerical representation. This faux-naïf attitude allows neuroimagers to manipulate two worlds simultaneously: by adhering to mathematization they gain scientific credence; while by exorbitantly processing colorful images they mesmerize non-professionals. As a corollary, neuroimagers gain support on all fronts (Beaulieu 2002). The persuasive power of NIs is abused as a communication tool to promote specific scientific ideology at the expense of rivaling theories (De Vos 2014). This combination of treacherous images, a cognitive bias toward visual imagery, and a conflicted scientific culture hidden from the public eye can quickly turn NIs to an ethically loaded gun.

An Alternative Account of Cognition: Emergence and Evolution

It seems incumbent to conclude the criticism offered here by proposing an alternative account of the story behind the brain-mind connection. To that end, I return to the critique of localization and ponder that while there is no doubt that reductionism has served science extremely well, does this success necessarily translate into a monopoly over fundamental ontology? Is reduction the best tool now that we know the brain is a complex, multi-player, multi-layered, spatiotemporally spread, non-linear, and heavily context-dependent system (e.g. (Ellis 2009))? What could be a more philosophically as well as physiologically apt interpretation of the link between the brain and the mind? To accept both that there are physiological underpinnings of cognition, emotions, and behavior, as well as that those phenomena exist only at the organismal level and above, forces us to identify and characterize a process by which electrochemical signal transduction is transformed into abstract, intangible, and vaporous thoughts and feelings. Since we currently lack the knowledge to offer any particular property as such a Holy Grail psychophysical compiler, I offer here (not as novelty) as candidates neuronal emergent properties.

Emergent properties are novel traits of a system that result from unique spatiotemporal relational dynamics of the building blocks of lower organization levels of that system, which are irreducible to the principles governing those lower levels (Rueger 2000; Silberstein 2002; Newman 2011)[17]. In this fluid interplay, building blocks may constrain - but not determine - the attributes of higher-level traits, such that the phenomenon generated at each level obeys the rules of that level, not of lower, and that

17. Emergence is often seen as an opposition to reductionism on account of irreducibility (Delehanty 2005).

higher levels may exert regulatory power over lower levels but not vice versa (Newsome 2009).

In brains, the wide range of neuronal elements, along with their contextual spatiotemporal orchestration of connectivity patterns and interactions in a complex 3D architecture, combine to generate novel processes that should be considered an emergent property of the nervous system as it complexifies and goes up organization levels. To date, a host of emergent properties have been documented, covering many aspects of neuronal operation, such as cortical oscillations (Wang 2010; Whittington et al. 2011), cortical coordination dynamics (Bressler and Kelso 2001), spiking rate (De Sancristóbal et al. 2013), circadian rhythms (Muraro, Pírez, and Ceriani 2013), electrical properties of dendritic spines (Yuste 2013), pre/postsynaptic terminal structure (Emes and Grant 2012), formation of neuronal assemblies (Fingelkurts, Fingelkurts, and Kähkönen 2005), synchronization of neuronal assemblies (Lindsey et al. 1997), functional information segregation (Ma et al. 2014), synaptic current and phase locking (Deco et al. 2008), and cognition in general (McIntosh 2000).

The idea here is that, given that brain activity is no stranger to emergent properties, higher-level neuronal actions combine to create the necessary conditions for the emergence of a quasi neuronal-independent process that utilize the dynamic interactions with the body and the environment to create specific portions of what we call the mind.

If we were to accept emergence as an unspecified mechanism shaping brains and cognition, then a next step is to accept that the system in question had undergone an evolutionary change (a change can be ontogenic only, but then it is of no consequence to the species). By agreeing that en evolutionary account is appropriate here, we then need to turn our attention to possible evolutionary mechanisms at play. Bunge (Bunge 1977) argued that radical novelties emerge out of previously existing things, such that emergence and levels of organization are dynamic orders and are features of an evolutionary process. In biological terms, what Bunge referred to, later became known as exaptation (Gould and Vrba 1982): the process by which features that now enhance fitness but were not built by natural selection for their current role have appeared.

Exaptation has already found its way into neuroscientific thought via neuronal re-use hypotheses, which argue that the complexification and evolution of the brain takes place via preservation, extension, and combination of existing networks (e.g. (Sporns and Kötter 2004; Just and Varma 2007)). The most sophisticated neural re-use theory is the massive redeployment hypothesis (Anderson 2007). It argues that existing components, which serve a specific purpose, are recruited for new purposes and combined to support new capacities without disrupting their previous functionality. Thus, each element does

only one thing, and it continues to do it regardless of the networks and complexes it is part of. It is far beyond the scope of this paper to develop the evolutionary exaptationist model of the transition from brain to mind, and it is presented here only as an abstract appetizer of a prolegomenon, a probable non-reductive account of the link between nervous systems and cognition, standing in opposition to the materialist localization project defended by neuroimagers.[18]

Conclusion

I argued here that the use of functional neuroimages for the purpose of supporting localization claims embedding the mind and consciousness exclusively in well-defined brain regions is an unethical scientific endeavor. By adopting a four-prong critique, I challenged the validity of the basic tenets and practices of neuroimaging at the technological, methodological, and philosophical levels. This criticism highlights lacunae that result in five degrees of separation between the biological phenomenon of neural response to a cognitive task and the NIs that allegedly represent it. I then examined the psychological impact of NIs and learned that they are borne out of a scientific culture with a strong penchant for aesthetics, a bias kept hidden from the public eye. The resultant hypothesis is that cognitive neuroscience's use of functional NIs is unethical by knowingly allowing flawed conclusions to trickle down from labs to policy makers, mass media, and the public, thus skewing public understanding of the fundamental issue of the mind-brain connection and cementing an erroneous interpretation of this problem. I am of the opinion that a scientific culture that cultivates and celebrates aesthetics over scientific accuracy and reliability; that knowingly disseminates distorted data masked by the appeal of heavily engineered images should not be surprised by fiascos such as the 2008 US presidential elections fMRI scan study. To pretend to be outraged by it is to turn a hypocritical blind eye to the academic climate that facilitated it while working toward perpetuating the very same problems that constitute a hurdle in the quest for a candid scientific effort to understand the mind and its possible neuronal basis.

18. Evolutionary theory will also allow further development of the distinction between neurotype and cognitype, by distinguishing between two possible candidates for selection processes.

References

Aguirre, Geoffrey K. 2014. "Functional neuroimaging: Technical, logical, and social perspectives." *Hastings Center Report* 44 (S2): S8-S18. doi: 10.1002/hast.294.

Aharoni, Eyal, Gina M. Vincent, Carla L. Harenski, Vince D. Calhoun, Walter Sinnott-Armstrong, Michael S. Gazzaniga, and Kent A. Kiehl. 2013. "Neuroprediction of future rearrest." *Proceedings of the National Academy of Sciences of the United States of America* 110 (15): 6223-6228. doi: 10.1073/pnas.1219302110.

Ahn, Woo-Young, Kenneth T Kishida, Xiaosi Gu, Terry Lohrenz, Ann Harvey, John R Alford, Kevin B Smith, Gideon Yaffe, John R Hibbing, Peter Dayan, and P. Read Montague. 2014. "Nonpolitical images evoke neural predictors of political ideology." *Current Biology* 24 (22): 2693-2699. doi: 10.1016/j.cub.2014.09.050.

Alvarez-Linera, Juan. "3T MRI: Advances in brain imaging." *European Journal of Radiology* 67 (3): 415-426. doi: 10.1016/j.ejrad.2008.02.045.

Anderson, James, Ania Mizgalewicz, and Judy Illes. 2012. "Reviews of functional MRI: The ethical dimensions of methodological critique." *PLoS ONE* 7 (8): e42836. doi: 10.1371/journal.pone.0042836.

Anderson, Michael L. 2007. "Massive redeployment, exaptation, and the functional integration of cognitive operations." *Synthese* 159 (3): 329-345. doi: 10.1007/s11229-007-9233-2.

Angermeyer, Matthias C., Anita Holzinger, Mauro G. Carta, and Georg Schomerus. 2011. "Biogenetic explanations and public acceptance of mental illness: Systematic review of population studies." *British Journal of Psychiatry* 199 (5): 367-372. doi: 10.1192/bjp.bp.110.085563.

Aron, Arthur, Helen Fisher, Debra J. Mashek, Greg Strong, Haifang Li, and Lucy L. Brown. 2005. "Reward, motivation, and emotion systems associated with early-stage intense romantic love." *Journal of Neurophysiology* 94 (1): 327-337. doi: 10.1152/jn.00838.2004.

Aru, Jaan, Nikolai Axmacher, Anne T. A. Do Lam, Juergen Fell, Christian E. Elger, Wolf Singer, and Lucia Melloni. 2012. "Local category-specific gamma band responses in the visual cortex do not reflect conscious perception." *The Journal of Neuroscience* 32 (43): 14909-14914. doi: 10.1523/jneurosci.2051-12.2012.

Ball, Tonio, Johanna Derix, Johanna Wentlandt, Birgit Wieckhorst, Oliver Speck, Andreas Schulze-Bonhage, and Isabella Mutschler. 2009. "Anatomical specificity of functional amygdala imaging of responses to stimuli with positive and negative

emotional valence." *Journal of Neuroscience Methods* 180 (1): 57-70. doi: 10.1016/j.jneumeth.2009.02.022.

Bartels, Andreas, and Semir Zeki. 2000. "The neural basis of romantic love." *NeuroReport* 11 (17): 3829-3834.

———. 2004. "The neural correlates of maternal and romantic love." *NeuroImage* 21 (3): 1155-1166. doi: 10.1016/j.neuroimage.2003.11.003.

Bartolo, Angela, Francesca Benuzzi, Luca Nocetti, Patrizia Baraldi, and Paolo Nichelli. 2006. "Humor comprehension and appreciation: An fMRI study." *Journal of Cognitive Neuroscience* 18 (11): 1789-1798. doi: 10.1162/jocn.2006.18.11.1789.

Beaulieu, Anne. 2002. "Images are not the (only) truth: Brain mapping, visual knowledge, and iconoclasm." *Science, Technology, and Human Values* 27 (1): 53-86. doi: 10.1177/016224390202700103.

Beauregard, Mario. 2009. "Effect of mind on brain activity: Evidence from neuroimaging studies of psychotherapy and placebo effect." *Nordic Journal of Psychiatry* 63 (1): 5-16. doi: 10.1080/08039480802421182.

Beauregard, Mario, and Vincent Paquette. 2006. "Neural correlates of a mystical experience in Carmelite nuns." *Neuroscience Letters* 405 (3): 186-190. doi: 10.1016/j.neulet.2006.06.060.

Bechtel, William. 2002. "Decomposing the brain: A long term pursuit." *Brain and Mind* 3 (2): 229-242. doi: 10.1023/A:1019980423053.

———. 2004. "The epistemology of evidence in cognitive neuroscience." In *Philosophy and the life sciences: A reader*, edited by Robert A. Skipper Jr., Colin Allen, Rachel A. Ankeny, Carl F. Craver, Lindley Darden, Greg Mikkelson and Robert Richardson. Cambridge, Mass.: MIT Press.

Beck, Diane M. 2010. "The appeal of the brain in the popular press." *Perspectives on Psychological Science* 5 (6): 762-766. doi: 10.1177/1745691610388779.

Bélanger, Mireille, Igor Allaman, and Pierre J Magistretti. 2011. "Brain energy metabolism: Focus on astrocyte-neuron metabolic cooperation." *Cell Metabolism* 14 (6): 724-738. doi: 10.1016/j.cmet.2011.08.016.

Bell, Emily, and Eric Racine. 2009. "Enthusiasm for functional magnetic resonance imaging (fMRI) often overlooks its dependence on task selection and performance." *American Journal of Bioethics* 9 (1): 23-25. doi: 10.1080/15265160802617894.

Bennett, Craig M., Abigail A. Baird, Michael B. Miller, and George L. Wolford. 2011. "Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for proper multiple comparisons correction." *Journal of Serendipitous and Unexpected Results* 1 (1): 1-5.

Bennett, Craig M., and Michael B. Miller. 2010. "How reliable are the results from functional magnetic resonance imaging?" *Annals of the New York Academy of Sciences* 1191 (1): 133-155. doi: 10.1111/j.1749-6632.2010.05446.x.

Bennett, Craig M., George L. Wolford, and Michael B. Miller. 2009. "The principled control of false positives in neuroimaging." *Social Cognitive and Affective Neuroscience* 4:417-422. doi: 10.1093/scan/nsp053.

Bennett, Max, and Peter Michael Stephan Hacker. 2003. *Philosophical foundations of neuroscience*. Oxford: Wiley-Blackwell.

Bermúdez i Badia, Sergi, Ulysses Bernardet, and Paul F. M. J. Verschure. 2010. "Non-linear neuronal responses as an emergent property of afferent networks: A case study of the locust Lobula giant movement detector." *PLoS Computational Biology* 6 (3): e1000701. doi: 10.1371/journal.pcbi.1000701.

Bichot, Narcisse P., S. Chenchal Rao, and Jeffrey D. Schall. 2001. "Continuous processing in macaque frontal cortex during visual search." *Neuropsychologia* 39 (9): 972-982. doi: 10.1016/S0028-3932(01)00022-7.

Block, Ned. 1996. "What is functionalism? (Revised version)." In *The Encyclopedia of Philosophy Supplement*, edited by Donald M. Borchert, 775. Macillan Library Reference.

Bluhm, Robyn. 2013. "New research, old problems: Methodological and ethical issues in fMRI research examining sex/gender differences in emotion processing." *Neuroethics* 6 (2): 319-330. doi: 10.1007/s12152-011-9143-3.

Bogen, Jim. 2002. "Epistemological custard pies from functional brain imaging." *Philosophy of Science* 69 (3): S59-S71.

Brechman, Jean, Chul-Joo Lee, and Joseph N. Cappella. 2009. "Lost in translation? A comparison of cancer-genetics reporting in the press release and its subsequent coverage in the press." *Science Communication* 30 (4): 453-474. doi: 10.1177/1075547009332649.

Bressler, Steven L., and J. A. Scott Kelso. 2001. "Cortical coordination dynamics and cognition." *Trends in Cognitive Sciences* 5 (1): 26-36. doi: 10.1016/S1364-6613(00)01564-3.

Buchsbaum, Bradley R., and Mark D'Esposito. 2008. "The search for the phonological store: From loop to convolution." *Journal of Cognitive Neuroscience* 20 (5): 762-778. doi: 10.1162/jocn.2008.20501.

Buckner, Randy L. 2003. "The hemodynamic inverse problem: Making inferences about neural activity from measured MRI signals." *Proceedings of the National Academy of Sciences of the United States of America* 100 (5): 2177-2179. doi: 10.1073/pnas.0630492100.

Bunge, Mario. 1977. "Emergence and the mind." *Neuroscience* 2 (4): 501-509. doi: 10.1016/0306-4522(77)90047-1.

Bunzl, Martin, Stephen José Hanson, and Russell A. Poldrack. 2010. "An exchange about localism." In *Foundational issues in human brain mapping*, edited by Stephen José Hanson and Martin Bunzl, 49-54. Cambridge, Mass: MIT Press.

Burock, Marc A. 2009. "Over-interpreting functional neuroimages." doi: http://philsci-archive.pitt.edu/4900/.

Burri, Regula Valérie. 2012. "Visual rationalities: Towards a sociology of image." *Current Sociology* 60 (1): 45-60. doi: 10.1177/0011392111426647.

———. 2013. "Visual power in action: Digital images and the shaping of medical practices." *Science as Culture* 22 (3): 367-387. doi: 10.1080/09505431.2013.768223.

Burwood, Stephen. 2009. "Are we our brains?" *Philosophical Investigations* 32 (2): 113-133. doi: 10.1111/j.1467-9205.2008.01366.x.

Button, Katherine S., John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafo. 2013. "Power failure: why small sample size undermines the reliability of neuroscience." *Nature Reviews Neuroscience* 14 (5): 365-376. doi: 10.1038/nrn3475.

Buzsáki, György. 2010. "Neural syntax: Cell assemblies, synapsembles, and readers." *Neuron* 68 (3): 362-385. doi: 10.1016/j.neuron.2010.09.023.

Byrge, Lisa, Olaf Sporns, and Linda B. Smith. 2014. "Developmental process emerges from extended brain-body-behavior networks." *Trends in Cognitive Sciences* 18 (8): 395-403. doi: 10.1016/j.tics.2014.04.010.

Cabeza, Roberto, and Lars Nyberg. 2000. "Imaging cognition II: An empirical review of 275 PET and fMRI studies." *Journal of Cognitive Neuroscience* 12 (1): 1-47. doi: 10.1162/08989290051137585.

Caesar, Kirsten, Kirsten Thomsen, and Martin Lauritzen. 2003. "Dissociation of spikes, synaptic activity, and activity-dependent increments in rat cerebellar blood flow by tonic synaptic inhibition." *Proceedings of the National Academy of Sciences of the United States of America* 100 (26): 16000-16005. doi: 10.1073/pnas.2635195100.

Carp, Joshua. 2012a. "On the plurality of (methodological) worlds: Estimating the analytic flexibility of fMRI experiments." *Frontiers in Neuroscience* 6:Article 149. doi: 10.3389/fnins.2012.00149.

———. 2012b. "The secret lives of experiments: Methods reporting in the fMRI literature." *NeuroImage* 63 (1): 289-300. doi: 10.1016/j.neuroimage.2012.07.004.

Casini, Silvia. 2011. "Magnetic resonance imaging (MRI) as mirror and portrait: MRI configurations between science and the arts." *Configurations* 19 (1): 73-99.

Cassels, Alan, Merrilee A. Hughes, Carol Cole, Barbara Mintzes, Joel Lexchin, and James P. McCormack. 2003. "Drugs in the news: an analysis of Canadian newspaper coverage of new prescription drugs." *Canadian Medical Association Journal* 168 (9): 1133-1137.

Caulfield, Timothy. 2005. "Popular media, biotechnology and the "cycle of hype"." *Journal of Health Law and Policy* 5:213-233.

Chan, Yu-Chen, Tai-Li Chou, Hsueh-Chih Chen, and Keng-Chen Liang. 2012. "Segregating the comprehension and elaboration processing of verbal jokes: An fMRI study." *NeuroImage* 61 (4): 899-906. doi: 10.1016/j.neuroimage.2012.03.052.

Chancellor, Bree, and Anjan Chatterjee. 2011. "Brain branding: When neuroscience and commerce collide." *AJOB Neuroscience* 2 (4): 18-27. doi: 10.1080/21507740.2011.611123.

Charest, Ian, Rogier A. Kievit, Taylor W. Schmitz, Diana Deca, and Nikolaus Kriegeskorte. 2014. "Unique semantic space in the brain of each beholder predicts perceived similarity." *Proceedings of the National Academy of Sciences* 111 (40): 14565-14570. doi: 10.1073/pnas.1402594111.

Check, Erika. 2005. "Ethicists urge caution over emotive power of brain scans." *Nature* 435 (7040): 254-255. doi: 10.1038/435254a

Chiel, Hillel J., and Randall D. Beer. 1997. "The brain has a body: adaptive behavior emerges from interactions of nervous system, body and environment." *Trends in Neuroscience* 20 (12): 553-557. doi: 10.1016/S0166-2236(97)01149-1.

Cikara, M., A. C. Jenkins, N. Dufour, and R. Saxe. 2014. "Reduced self-referential neural response during intergroup competition predicts competitor harm." *NeuroImage* 96 (1): 36-43. doi: 10.1016/j.neuroimage.2014.03.080.

Cohn, Simon. 2010. "Picturing the brain inside, revealing the illness outside: A comparison of the different meanings attributed to brain scans by scientists and patients." In *Technologized images, technologized bodies: Anthropological approaches to a new politics of vision*, edited by Jeanette Edwards, Penelope Harvey and Peter Wade, 65-84. Oxford: Berghahn.

Cohn, Simon, Joseph Dumit, and Andreas Roepstorff. 2003. Neuroscience promises and the challenge of brain imaging to the conceptions of mental and physical illness. Economic and Social Research Council.

Coltheart, Max. 2004. "Brain imaging, connectionism and cognitive neuropsychology." *Cognitive Neuropsychology* 21 (1): 21-25. doi: 10.1080/02643290342000159.

———. 2006. "What has functional neuroimaging told us about the mind (so far)?" *Cortex* 42 (3): 323-331. doi: 10.1016/S0010-9452(08)70358-7.

Conrad, Peter. 1997. "Public eyes and private genes: Historical frames, news constructions, and social problems." *Social Problems* 44 (2): 139-154. doi: 10.1525/sp.1997.44.2.03x0219k.

Cox, Simon R, Karen J Ferguson, Natalie A Royle, Susan D Shenkin, Sarah E MacPherson, Alasdair M. J. MacLullich, Ian J Deary, and Joanna M Wardlaw. 2014. "A systematic review of brain frontal lobe parcellation techniques in magnetic resonance imaging." *Brain Structure and Function* 219 (1): 1-22. doi: 10.1007/s00429-013-0527-5.

Craver, Carl F. 2005. "Beyond reduction: Mechanisms, multifield integration, and the unity of science." *Studies in History and Philosophy of Biological and Biomedical Sciences* 36:373-396. doi: 10.1016/j.shpsc.2005.03.008.

Crick, Francis. 1996. "Visual perception: rivalry and consciousness." *Nature* 379 (6565): 485-486. doi: 10.1038/379485a0.

Damoiseaux, Jessica S., and Michael D. Greicius. 2009. "Greater than the sum of its parts: a review of studies combining structural connectivity and resting-state functional connectivity." *Brain Structure and Function* 213 (6): 525-533. doi: 10.1007/s00429-009-0208-6.

Dar-Nimrod, Ilan, and Steven J. Heine. 2011. "Genetic essentialism: On the deceptive determinism of DNA." *Psychological Bulletin* 137 (5): 800-818. doi: 10.1037/a0021860

David, Sean P., Jennifer J. Ware, Isabella M. Chu, Pooja D. Loftus, Paolo Fusar-Poli, Joaquim Radua, Marcus R. Munafò, and John P. A. Ioannidis. 2013. "Potential reporting bias in fMRI studies of the brain." *PLoS ONE* 8 (7): e70104. doi: 10.1371/journal.pone.0070104.

de Graaf, Tom A., Po-Jang Hsieh, and Alexander T. Sack. 2012. "The 'correlates' in neural correlates of consciousness." *Neuroscience and Biobehavioral Reviews* 36 (1): 191-197. doi: 10.1016/j.neubiorev.2011.05.012.

De Sancristóbal, Belén, Raul Vicente, Jose Maria Sancho, and Jordi Garcia-Ojalvo. 2013. "Emergent bimodal firing patterns implement different encoding strategies during gamma-band oscillations." *Frontiers in Computational Neuroscience* 7. doi: 10.3389/fncom.2013.00018.

De Vos, Jan. 2014. "The iconographic brain. A critical philosophical inquiry into (the resistance of) the image." *Frontiers in Human Neuroscience* 8. doi: 10.3389/fnhum.2014.00300.

Deco, Gustavo, Viktor K. Jirsa, Peter A. Robinson, Michael Breakspear, and Karl Friston. 2008. "The dynamic brain: From spiking neurons to neural masses and cortical fields." *PLoS Computational Biology* 4 (8): e1000092. doi: 10.1371/journal.pcbi.1000092.

Devlin, Joseph T., Richard P. Russell, Matt H. Davis, Cathy J. Price, James Wilson, Helen E. Moss, Paul M. Matthews, and Lorraine K. Tyler. 2000. "Susceptibility-induced loss of signal: Comparing PET and fMRI on a semantic task." *NeuroImage* 11 (6): 589-600. doi: 10.1006/nimg.2000.0595.

Devonshire, Ian M., Nikos G. Papadakis, Michael Port, Jason Berwick, Aneurin J. Kennerley, John E. W. Mayhew, and Paul G. Overton. 2012. "Neurovascular coupling is brain region-dependent." *NeuroImage* 59 (3): 1997-2006. doi: 10.1016/j.neuroimage.2011.09.050.

Devor, Anna, Elizabeth M. C. Hillman, Peifang Tian, Christian Waeber, Ivan C. Teng, Lana Ruvinskaya, Mark H. Shalinsky, Haihao Zhu, Robert H. Haslinger, Suresh N. Narayanan, Istvan Ulbert, Andrew K. Dunn, Eng H. Lo, Bruce R. Rosen, Anders M. Dale, David Kleinfeld, and David A. Boas. 2008. "Stimulus-induced changes in blood flow and 2-deoxyglucose uptake dissociate in ipsilateral somatosensory cortex." *The Journal of Neuroscience* 28 (53): 14347-14357. doi: 10.1523/jneurosci.4307-08.2008.

Diederen, Kelly M. J., L. Charbonnier, Sebastiaan F. W. Neggers, Remko van Lutterveld, Kirstin Daalman, C. W. Slotema, R. S. Kahn, and Iris E. C. Sommer.

2013. "Reproducibility of brain activation during auditory verbal hallucinations." *Schizophrenia Research* 146 (1): 320-325. doi: 10.1016/j.schres.2013.01.025.

Dingfelder, Sadie F. 2008. "Do psychologists have "neuron envy"?" *Monitor on Psychology* 39 (6): 26-27.

Dolcos, Florin, Ekaterina Denkova, and Sanda Dolcos. 2013. "Neural correlates of emotional memories: A review of evidence from brain imaging studies." *Psychologia* 55 (2): 80-111. doi: 10.2117/psysoc.2012.80.

Downie, Jocelyn, and Matthew Herder. 2007. "Reflections on the commercialization of research conducted in public institutions in Canada." *McGill Health Law Publication* 1 (1): 23-44.

Dumit, Joseph. 2004. *Picturing personhood: Brain scans and biomedical Identity*. Edited by Paul Rabinow, *In-Formation*. Princeton, NJ: Princeton University Press.

Duncan, John. 2001. "An adaptive coding model of neural function in prefrontal cortex." *Nature Reviews Neuroscience* 2 (11): 820-829. doi: 10.1038/35097575.

Easterbrook, Philippa J., Ramana Gopalan, Jesse A. Berlin, and David R. Matthews. 1991. "Publication bias in clinical research." *The Lancet* 337 (8746): 867-872. doi: 10.1016/0140-6736(91)90201-Y.

Edelman, Gerald M., and Joseph A. Gally. 2001. "Degeneracy and complexity in biological systems." *Proceedings of the National Academy of Sciences of the United States of America* 98 (24): 13763-13768. doi: 10.1073/pnas.231499798.

Ekstron, Arne. 2010. "How and when the fMRI BOLD signal relates to underlying neural activity: The danger in dissociation." *Brain Research Reviews* 62 (2): 233-244. doi: 10.1016/j.brainresrev.2009.12.004.

Ellis, George F. R. 2009. "Top-down causation and the human brain." In *Downward causation and the neurobiology of free will*, edited by Nancey Murphy, George F. R. Ellis and Timothy O'Connor, 63-81. Berlin: Springer.

Emes, Richard D., and Seth G. N. Grant. 2012. "Evolution of synapse complexity and diversity." *Annual Review of Neuroscience* 35:111-131. doi: 10.1146/annurev-neuro-062111-150433.

Escartin, Carole, and Nathalie Rouach. 2013. "Astroglial networking contributes to neurometabolic coupling." *Frontiers in Neuroenergetics* 5. doi: 10.3389/fnene.2013.00004.

Fanelli, Daniele. 2010. ""Positive" results increase down the hierarchy of the sciences." *PLoS ONE* 5 (4): e10068. doi: 10.1371/journal.pone.0010068.

Fava, Giovanni A. 2006. "The intellectual crisis of psychiatric eesearch." *Psychotherapy and Psychosomatics* 75 (4): 202-208. doi: 10.1159/000092890.

Fellows, Lesley K., Andrea S. Heberlein, Dawn A. Morales, Geeta Shivde, Sara Waller, and Denise H. Wu. 2005. "Method matters: An empirical study of impact in cognitive neuroscience." *Journal of Cognitive Neuroscience* 17 (6): 850-858. doi: 10.1162/0898929054021139.

Fernandez-Duque, Diego, Jessica Evans, Colton Christian, and Sara D. Hodges. 2014. "Superfluous neuroscience information makes explanations of psychological phenomena more appealing." *Journal of Cognitive Neuroscience*:1-19. doi: 10.1162/jocn_a_00750.

Fiedler, Klaus. 2011. "Voodoo correlations are everywhere - Not only in neuroscience." *Perspectives on Psychological Science* 6 (2): 163-171. doi: 10.1177/1745691611400237.

Figdor, Carrie. 2010. "Neuroscience and the multiple realization of cognitive functions." *Philosophy of Science* 77 (3): 419-456.

Figley, Chase R., and Patrick W. Stroman. 2011. "The role(s) of astrocytes and astrocyte activity in neurometabolism, neurovascular coupling, and the production of functional neuroimaging signals." *European Journal of Neuroscience* 33 (4): 577-588. doi: 10.1111/j.1460-9568.2010.07584.x.

Fingelkurts, Andrew A., Alexander A. Fingelkurts, and Seppo Kähkönen. 2005. "Functional connectivity in the brain - Is it an elusive concept?" *Neuroscience and Biobehavioral Reviews* 28 (8): 827-836. doi: 10.1016/j.neubiorev.2004.10.009.

Fishell, Gord, and Nathaniel Heintz. 2013. "The neuron identity problem: Form meets function." *Neuron* 80 (3): 602-612. doi: 10.1016/j.neuron.2013.10.035.

Flusser, Vilém. 2000. *Towards a philosophy of photography*. Translated by Anthony Mathews. London: Reaktion books. Original edition, 1983.

Fox, Kieran C. R., R. Nathan Spreng, Melissa Ellamil, Jessica R. Andrews-Hanna, and Kalina Christoff. 2015. "The wandering brain: Meta-analysis of functional neuroimaging studies of mind-wandering and related spontaneous thought processes." *NeuroImage* (0). doi: 10.1016/j.neuroimage.2015.02.039.

Fox, Michael D., and Marcus E. Raichle. 2007. "Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging." *Nature Reviews Neuroscience* 8 (9): 700-711.

Fox, Michael D., Abraham Z. Snyder, Justin L. Vincent, Maurizio Corbetta, David C. Van Essen, and Marcus E. Raichle. 2005. "The human brain is intrinsically organized into dynamic, anticorrelated functional networks." *Proceedings of the National Academy of Sciences of the United States of America* 102 (27): 9673-9678. doi: 10.1073/pnas.0504136102.

Fox, Peter T., and Karl J. Friston. 2012. "Distributed processing; distributed functions?" *NeuroImage* 61 (2): 407-426. doi: 10.1016/j.neuroimage.2011.12.051.

Francis, Gregory, Jay Tanzman, and William J. Matthews. 2014. "Excess success for psychology articles in the journal ‹italic›Science‹/italic›." *PLoS ONE* 9 (12): e114255. doi: 10.1371/journal.pone.0114255.

Fransson, Peter. 2005. "Spontaneous low-frequency BOLD signal fluctuations: An fMRI investigation of the resting-state default mode of brain function hypothesis." *Human Brain Mapping* 26 (1): 15-29. doi: 10.1002/hbm.20113.

Friston, Karl J., Pia Rotshtein, Joy J. Geng, Philipp Sterzer, and Rik N. Henson. 2010. "A critique of functional localizers." In *Foundational issues in human brain mapping*, edited by Stephen José Hanson and Martin Bunzl, 3-24. Cambridge, Mass.: MIT Press.

Frow, Emma K. 2012. "Drawing a line: Setting guidelines for digital image processing in scientific journal articles." *Social Studies of Science* 42 (3): 369-392. doi: 10.1177/0306312712444303.

Fusar-Poli, Paolo, and M. R. Broome. 2007. "Love and brain: From mereological fallacy to "folk" neuroimaging." *Psychiatry Research: Neuroimaging* 154 (3): 285-286. doi: 10.1016/j.pscychresns.2006.11.001.

Gabrieli, John D E., Satrajit S Ghosh, and Susan Whitfield-Gabrieli. 2015. "Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience." *Neuron* 85 (1): 11-26. doi: 10.1016/j.neuron.2014.10.047.

García-Eulate, Reyes, David García-García, Pablo D. Dominguez, Jose J. Noguera, Esther De Luis, María C. Rodriguez-Oroz, and Jose L. Zubieta. 2011. "Functional bold MRI: Advantages of the 3 T vs. the 1.5 T." *Clinical Imaging* 35 (3): 236-241. doi: 10.1016/j.clinimag.2010.07.003.

Geissler, Alexander, Rupert Lanzenberger, Markus Barth, Amir Reza Tahamtan, Denny Milakara, Andreas Gartus, and Roland Beisteiner. 2005. "Influence of fMRI smoothing procedures on replicability of fine scale motor localization." *NeuroImage* 24 (2): 323-331. doi: 10.1016/j.neuroimage.2004.08.042.

Gerlach, Christian. 2007. "A review of functional imaging studies on category specificity." *Journal of Cognitive Neuroscience* 19 (2): 296-314. doi: 10.1162/jocn.2007.19.2.296.

Glannon, Walter. 2009. "Our brains are not us." *Bioethics* 23 (6): 321-3229. doi: 10.1111/j.1467-8519.2009.01727.x.

Goense, Jozien, Kevin Whittingstall, and Nikos K. Logothetis. 2012. "Neural and BOLD responses across the brain." *Wiley Interdisciplinary Reviews: Cognitive Science* 3 (1): 75-86. doi: 10.1002/wcs.153.

Gonon, Francois, Erwan Bezard, and Thomas Boraud. 2011. "Misrepresentation of neuroscience data might give rise to misleading conclusions in the media: The case of attention deficit hyperactivity disorder." *PLoS ONE* 6 (1): e14618. doi: 10.1371/journal.pone.0014618.

Gonzalez-Castillo, Javier, Ziad S. Saad, Daniel A. Handwerker, Souheil J. Inati, Noah Brenowitz, and Peter A. Bandettini. 2012. "Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis." *Proceedings of the National Academy of Sciences of the United States of America* 109 (14): 5487-5492. doi: 10.1073/pnas.1121049109.

Goodman, Nelson. 1976. *Languages of art: An approach to the theory of symbols*. 2nd ed. Indianapolis, IN: Hackett.

Gould, Stephen Jay, and Elisabeth S. Vrba. 1982. "Exaptation: a missing term in the science of form." *Paleobiology* 8 (1): 4-15.

Gross, Charles G. 2002. "The genealogy of the "Grandmother cell"." *The Neuroscientist* 8 (5): 512-518. doi: 10.1177/107385802237175.

Gruber, David, and Jacob A. Dickerson. 2012. "Persuasive images in popular science: Testing judgments of scientific reasoning and credibility." *Public Understanding of Science* 21 (8): 938-948. doi: 10.1177/0963662512454072.

Guo, Qing, Lehana Thabane, Geoffrey Hall, Margaret McKinnon, Ron Goeree, and Eleanor Pullenayegum. 2014. "A systematic review of the reporting of sample size calculations and corresponding data components in observational functional magnetic resonance imaging studies." *NeuroImage* 86:172-181. doi: 10.1016/j.neuroimage.2013.08.012.

Han, Shihui, and Yina Ma. 2014. "Cultural differences in human brain activity: A quantitative meta-analysis." *NeuroImage* 99 (0): 293-300. doi: 10.1016/j.neuroimage.2014.05.062.

Hardcastle, Valerie Gray, and C. Matthew Stewart. 2002. "What do brain data really show?" *Philosophy of Science* 69 (3): 572-582. doi: 10.1086/341769.

Harley, Trevor A. 2004. "Does cognitive neuropsychology have a future?" *Cognitive Neuropsychology* 21 (1): 3-16. doi: 10.1080/02643290342000131.

Harris, Sam, Jonas T. Kaplan, Ashley Curiel, Susan Y. Bookheimer, Marco Iacoboni, and Mark S. Cohen. 2009. "The neural correlates of religious and nonreligious belief." *PLoS ONE* 4 (10): e7272. doi: 10.1371/journal.pone.0007272.

Harris, Sam, Sameer A. Sheth, and Mark S. Cohen. 2008. "Functional neuroimaging of belief, disbelief, and uncertainty." *Annals of Neurology* 63 (2): 141-147. doi: 10.1002/ana.21301.

Haslam, Nick. 2011. "Genetic essentialism, neuroessentialism, and stigma: Commentary on Dar-Nimrod and Heine (2011)." *Psychological Bulletin* 137 (5): 819-824. doi: 10.1037/a0022386.

Haueis, Philipp. 2012. "The fuzzy brain. Vagueness and mapping connectivity of the human cerebral cortex." *Frontiers in Neuroanatomy* 6:37. doi: 10.3389/fnana.2012.00037.

Haxby, James V. 2010. "Multivariate pattern analysis of fMRI data: High-dimensional spaces for neural and cognitive representations." In *Foundational issues in human brain mapping*, edited by Stephen José Hanson and Martin Bunzl, 55-68. Cambridge, Mass.: MIT Press.

Haxby, James V., Susan M. Courtney, and Vincent P. Clark. 1998. "Functional magnetic resonance imaging and the study of attention." In *The attentive brain*, edited by Raja Parasuraman, 123-142. Cambridge, Mass.: MIT Press.

Haxby, James V., M. Ida Gobbini, Maura L. Furey, Alumit Ishai, Jennifer L. Schouten, and Pietro Pietrini. 2001. "Distributed and overlapping representations of faces and objects in ventral temporal cortex." *Science* 293 (5539): 2425-2430. doi: 10.1126/science.1063736.

Henny, Pablo, Matthew T. C. Brown, Augustus Northrop, Macarena Faunes, Mark A. Ungless, Peter J. Magill, and J. Paul Bolam. 2012. "Structural correlates of heterogeneous in vivo activity of midbrain dopaminergic neurons." *Nature Neuroscience* 15 (4): 613-619. doi: 10.1038/nn.3048.

Henson, Richard. 2005. "What can functional neuroimaging tell the experimental psychologist?" *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology* 58 (2): 193-233. doi: 10.1080/02724980443000502.

Hermes, Dora, Kai J. Miller, Mariska J. Vansteensel, Erik J. Aarnoutse, Frans S. S. Leijten, and Nick F. Ramsey. 2012. "Neurophysiologic correlates of fMRI in human motor cortex." *Human Brain Mapping* 33 (7): 1689-1699. doi: 10.1002/hbm.21314.

Hoenig, Klaus, Christiane K. Kuhl, and Lukas Scheef. 2005. "Functional 3.0-T MR assessment of higher cognitive function: are there advantages over 1.5-T imaging?" *Radiology* 234 (3): 860-868. doi: doi:10.1148/radiol.2343031565.

Hong, Wei, and John P. Walsh. 2009. "For money or glory? Commercialization, competition and secrecy in the entrepreneurial university." *Sociological Quarterly* 50 (1): 145-171. doi: 10.1111/j.1533-8525.2008.01136.x.

Hook, Cayce J., and Martha J. Farah. 2013. "Look again: Effects of brain images and mind-brain dualism on lay evaluations of research." *Journal of Cognitive Neuroscience* 25 (9): 1397-1405. doi: 10.1162/jocn_a_00407.

Huber, Christian G., and Johannes Huber. 2009. "Epistemological considerations on neuroimaging - A crucial prerequisite for neuroethics." *Bioethics* 23 (6): 340-348. doi: 10.1111/j.1467-8519.2009.01728.x.

Huber, Lara. 2009. "Imaging the brain: Visualising "pathological entities"? Searching for reliable protocols within psychiatry and their impact on the understanding of psychiatric diseases." *Poiesis und Praxis* 6 (2): 27-41. doi: 10.1007/s10202-008-0055-1.

Huo, Bing-Xing, Jared B. Smith, and Patrick J. Drew. 2014. "Neurovascular coupling and decoupling in the cortex during voluntary locomotion." *The Journal of Neuroscience* 34 (33): 10975-10981. doi: 10.1523/jneurosci.1369-14.2014.

Iadecola, Costantino. 2004. "Neurovascular regulation in the normal brain and in Alzheimer's disease." *Nature Reviews Neuroscience* 5 (5): 347-360. doi: 10.1038/nrn1387.

Ihnen, Sarah K. Z., Jessica A. Church, Steven E. Petersen, and Bradley L. Schlaggar. 2009. "Lack of generalizability of sex differences in the fMRI BOLD activity associated with language processing in adults." *NeuroImage* 45 (3): 1020-1032. doi: 10.1016/j.neuroimage.2008.12.034.

Ikeda, Kenji, Shinji Kitagami, Tomoyo Takahashi, Yosuke Hattori, and Yuichi Ito. 2013. "Neuroscientific information bias in metacomprehension: The effect of brain images

on metacomprehension judgment of neuroscience research." *Psychonomic Bulletin & Review* 20 (6): 1357-1363. doi: 10.3758/s13423-013-0457-5.

Ingre, Michael. 2013. "Why small low-powered studies are worse than large high-powered studies and how to protect against "trivial" findings in research: Comment on Friston (2012)." *NeuroImage* 81 (0): 496-498. doi: 10.1016/j.neuroimage.2013.03.030.

Ioannidis, John P. A. 2005a. "Contradicted and initially stronger effects in highly cited clinical research." *JAMA* 294 (2): 218-228. doi: 10.1001/jama.294.2.218.

———. 2005b. "Why most published research findings are false." *PLoS Medicine* 2 (8): e124. doi: 10.1371/journal.pmed.0020124.

Ioannidis, John P. A., Marcus R. Munafo, Paolo Fusar-Poli, Brian A. Nosek, and Sean P. David. 2014. "Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention." *Trends in Cognitive Sciences* 18 (5): 235-241. doi: 10.1016/j.tics.2014.02.010.

Jbabdi, Saad, Stamatios N. Sotiropoulos, and Timothy E. Behrens. 2013. "The topographic connectome." *Current Opinion in Neurobiology* 23 (2): 207-215. doi: 10.1016/j.conb.2012.12.004.

John, Leslie K., George Loewenstein, and Drazen Prelec. 2012. "Measuring the prevalence of questionable research practices with incentives for truth telling." *Psychological Science* 23 (5): 524-532. doi: 10.1177/0956797611430953.

Joyce, KarenE, and Satoru Hayasaka. 2012. "Development of PowerMap: A software package for statistical power calculation in neuroimaging studies." *Neuroinformatics* 10 (4): 351-365. doi: 10.1007/s12021-012-9152-3.

Joyce, Kelly. 2005. "Appealing images: Magnetic resonance imaging and the production of authoritative knowledge." *Social Studies of Science* 35 (3): 437-462. doi: 10.1177/0306312705050180.

———. 2008. *Magnetic appeal: MRI and the myth of transparency*. Ithaca, NY: Cornell University Press

Jukovskaya, Natalya, Pascale Tiret, Jérôme Lecoq, and Serge Charpak. 2011. "What does local functional hyperemia tell about local neuronal activation?" *The Journal of Neuroscience* 31 (5): 1579-1582. doi: 10.1523/jneurosci.3146-10.2011.

Just, Marcel Adam, and Sashank Varma. 2007. "The organization of thinking: What functional brain imaging reveals about the neuroarchitecture of complex cognition."

Cognitive, Affective & Behavioral Neuroscience 7 (3): 153-191. doi: 10.3758/CABN.7.3.153.

Kapogiannis, Dimitrios, Aron K. Barbey, Michael Su, Giovanna Zamboni, Frank Krueger, and Jordan Grafman. 2009. "Cognitive and neural foundations of religious belief." *Proceedings of the National Academy of Sciences of the United States of America* 106 (12): 4876-4881. doi: 10.1073/pnas.0811717106.

Kaposy, Chris. 2008. "Ethical muscle and scientific interests: A role for philosophy in scientific research." *Quarterly Review of Biology* 83 (1): 77-86. doi: 10.1086/529565.

Kawabata Duncan, Keith J., and Joseph T. Devlin. 2011. "Improving the reliability of functional localizers." *NeuroImage* 57 (3): 1022-1030. doi: 10.1016/j.neuroimage.2011.05.009.

Keehner, Madeleine, and Martin H. Fischer. 2011. "Naive realism in public perceptions of neuroimages." *Nature Reviews Neuroscience* 12 (2): 118-65. doi: 10.1038/nrn2773-c1.

Kim, Jaegwon. 2006. "Emergence: Core ideas and issues." *Synthese* 151 (3): 547-559. doi: 10.1007/s11229-006-9025-0.

Klein, Colin. 2012. "Cognitive ontology and region- versus network-oriented analyses." *Philosophy of Science* 79 (5): 952-960. doi: 10.1086/667843.

Kong, Feng, Siyuan Hu, Xu Wang, Yiying Song, and Jia Liu. 2015. "Neural correlates of the happy life: The amplitude of spontaneous low frequency fluctuations predicts subjective well-being." *NeuroImage* 107:136-145. doi: 10.1016/j.neuroimage.2014.11.033.

Koren, Gideon, and Naomi Klein. 1991. "Bias against negative studies in newspaper reports of medical research." *Journal of the American Medical Association* 266 (13): 1824-1826. doi: 10.1001/jama.266.13.1824.

Koster-Hale, Jorie, Rebecca Saxe, James Dungan, and Liane L. Young. 2013. "Decoding moral judgments from neural representations of intentions." *Proceedings of the National Academy of Sciences of the United States of America* 110 (14): 5648-5653. doi: 10.1073/pnas.1207992110.

Kraff, Oliver, Anja Fischer, Armin M. Nagel, Christoph Mönninghoff, and Mark E. Ladd. 2015. "MRI at 7 tesla and above: Demonstrated and potential capabilities." *Journal of Magnetic Resonance Imaging* 41 (1): 13-33. doi: 10.1002/jmri.24573.

Kriegeskorte, Nikolaus, W Kyle Simmons, Patrick S. F. Bellgowan, and Chris I. Baker. 2009. "Circular analysis in systems neuroscience: The dangers of double dipping." *Nature Neuroscience* 12 (5): 535-540. doi: 10.1038/nn.2303.

Kringelbach, Morten L., and Edmund T. Rolls. 2004. "The functional neuroanatomy of the human orbitofrontal cortex: Evidence from neuroimaging and neuropsychology." *Progress in Neurobiology* 72 (5): 341-372. doi: 10.1016/j.pneurobio.2004.03.006.

Kvaale, Erlend P, William H Gottdiener, and Nick Haslam. 2013. "Biogenetic explanations and stigma: A meta-analytic review of associations among laypeople." *Social Science and Medicine* 96:95-103. doi: 10.1016/j.socscimed.2013.07.017.

Laufs, H., K. Krakow, P. Sterzer, E. Eger, A. Beyerle, A. Salek-Haddadi, and A. Kleinschmidt. 2003. "Electroencephalographic signatures of attentional and cognitive default modes in spontaneous brain activity fluctuations at rest." *Proceedings of the National Academy of Sciences of the United States of America* 100 (19): 11053-11058. doi: 10.1073/pnas.1831638100.

Lavazza, Andrea, and Mario De Caro. 2010. "Not so fast. On some bold neuroscientific claims concerning human agency." *Neuroethics* 3 (1): 23-41. doi: 10.1007/s12152-009-9053-9.

Lebowitz, Matthew S., and Woo-kyoung Ahn. 2014. "Effects of biological explanations for mental disorders on clinicians' empathy." *Proceedings of the National Academy of Sciences of the United States of America* 111 (50): 17786-17790. doi: 10.1073/pnas.1414058111.

Lee, Kevin F. H., Cary Soares, and Jean-Claude Beique. 2012. "Examining form and function of dendritic spines." *Neural Plasticity*:Article 704103. doi: 10.1155/2012/704103.

Lindsey, Bruce G., Kendall F. Morris, Roger Shannon, and G. L. Gerstein. 1997. "Repeated patterns of distributed synchrony in neuronal assemblies." *Journal of Neurophysiology* 78 (3): 1714-1719.

Logothetis, Nikos K. 2003. "The neural basis of the blood-oxygen-leveldependent functional magnetic resonance imaging signal." In *The physiology of cognitive processes*, edited by A. Parker, A. Derrington and C. Blakemore, 62-116. New York: Oxford University Press.

———. 2008. "What we can do and what we cannot do with fMRI." *Nature* 453 (7197): 869-878. doi: 10.1038/nature06976.

Logothetis, Nikos K., Jon Pauls, Mark Augath, Torsten Trinath, and Axel Oeltermann. 2001. "Neurophysiological investigation of the basis of the fMRI signal." *Nature* 412 (6843): 150-157. doi: 10.1038/35084005.

Loosemore, Richard, and Trevor A. Harley. 2010. "Brains and minds: On the usefulness of localization data to cognitive psychology." In *Foundational issues in human brain mapping*, edited by Stephen José Hanson and Martin Bunzl, 217-240. Cambridge, Mass: MIT Press.

Lynch, Michael. 1991. "Science in the age of mechanical reproduction: Moral and epistemic relations between diagrams and photographs." *Biology and Philosophy* 6:205-226. doi: 10.1007/BF02426838.

Ma, Liya, James M. Hyman, Adrian J. Lindsay, Anthony G. Phillips, and Jeremy K. Seamans. 2014. "Differences in the emergent coding properties of cortical and striatal ensembles." *Nature Neuroscience* 17 (8): 1100-1106. doi: 10.1038/nn.3753.

Martell, Daniel A. 2009. "Neuroscience and the law: Philosophical differences and practical constraints." *Behavioral Sciences and the Law* 27 (2): 123-136. doi: 10.1002/bsl.853.

Mayhew, Stephen D., Karen J. Mullinger, Andrew P. Bagshaw, Richard Bowtell, and Susan T. Francis. 2014. "Investigating intrinsic connectivity networks using simultaneous BOLD and CBF measurements." *NeuroImage* 99:111-121. doi: 10.1016/j.neuroimage.2014.05.042.

McCabe, David P. , and Alan D. Castel. 2008. "Seeing is believing: The effect of brain images on judgments of scientific reasoning." *Cognition* 107 (1): 343-352. doi: 10.1016/j.cognition.2007.07.017.

McGonigle, D. J., A. M. Howseman, B. S. Athwal, K. J. Friston, R. S. J. Frackowiak, and A. P. Holmes. 2000. "Variability in fMRI: An Examination of Intersession Differences." *NeuroImage* 11 (6): 708-734. doi: 10.1006/nimg.2000.0562.

McIntosh, Anthony Randal. 2000. "Towards a network theory of cognition." *Neural Networks* 13 (8-9): 861-870. doi: 10.1016/S0893-6080(00)00059-9.

Meinertzhagen, Ian A. , Shin-ya Takemura, Zhiyuan Lu, Songling Huang, Shuying Gao, Chun-Yuan Ting, and Chi-Hon Lee. 2009. "From form to function: The ways to know a neuron." *Journal of Neurogenetics* 23 (1-2): 68-77. doi: 10.1080/01677060802610604.

Melchers, Martin, Sebastian Markett, Christian Montag, Peter Trautner, Bernd Weber, Bernd Lachmann, Pauline Buss, Rebekka Heinen, and Martin Reuter. 2015. "Reality

TV and vicarious embarrassment: An fMRI study." *NeuroImage* (0). doi: 10.1016/j.neuroimage.2015.01.022.

Meurk, Carla, Adrian Carter, Wayne Hall, and Jayne Lucke. 2014. "Public understandings of addiction: Where do neurobiological explanations fit?" *Neuroethics* 7 (1): 51-62. doi: 10.1007/s12152-013-9180-1.

Michael, Robert B., Eryn J. Newman, Matti Vuorre, Geoff Cumming, and Maryanne Garry. 2013. "On the (non)persuasive power of a brain image." *Psychonomic Bulletin & Review* 20 (4): 720-725. doi: 10.3758/s13423-013-0391-6.

Miller, Jeff, and Steven A. Hackley. 1992. "Electrophysiological evidence for temporal overlap among contingent mental processes." *Journal of Experimental Psychology: General* 121 (2): 195-209. doi: 10.1037/0096-3445.121.2.195.

Miller, Michael B, and John Darrell Van Horn. 2007. "Individual variability in brain activations associated with episodic retrieval: A role for large-scale databases." *International Journal of Psychophysiology* 63 (2): 205-213. doi: 10.1016/j.ijpsycho.2006.03.019.

Miller, Michael B., Christa-Lynn Donovan, Craig M. Bennett, Elissa M. Aminoff, and Richard E. Mayer. 2012. "Individual differences in cognitive style and strategy predict similarities in the patterns of brain activity between individuals." *NeuroImage* 59 (1): 83-93. doi: 10.1016/j.neuroimage.2011.05.060.

Miller, Michael B., Christa-Lynn Donovan, John Darrell Van Horn, Elaine German, Peter Sokol-Hessner, and George L. Wolford. 2009. "Unique and persistent individual patterns of brain activity across different memory retrieval tasks." *NeuroImage* 48 (3): 625-635. doi: 10.1016/j.neuroimage.2009.06.033.

Mishra, Asht Mangal, Damien J. Ellens, Ulrich Schridde, Joshua E. Motelow, Michael J. Purcaro, Matthew N. DeSalvo, Miro Enev, Basavaraju G. Sanganahalli, Fahmeed Hyder, and Hal Blumenfeld. 2011. "Where fMRI and electrophysiology agree to disagree: Corticothalamic and striatal activity patterns in the WAG/Rij rat." *The Journal of Neuroscience* 31 (42): 15053-15064. doi: 10.1523/jneurosci.0101-11.2011.

Mole, Christopher, Corey Kubatzky, Jan Plate, Rawdon Waller, Marilee Dobbs, and Marc Nardone. 2007. "Faces and brains: The limitations of brain scanning in cognitive science." *Philosophical Psychology* 20 (2): 197-207. doi: 10.1080/09515080701209380.

Monti, Martin M. 2011. "Statistical analysis of fMRI time-series: A critical review of the GLM approach." *Frontiers in Human Neuroscience* 5:28. doi: 10.3389/fnhum.2011.00028.

Mundale, Jennifer. 2002. "Concepts of localization: Balkanization in the brain." *Brain and Mind* 3 (3): 313-330. doi: 10.1023/A:1022912227833.

Muraro, Nara I., Nicolás Pírez, and María Fernanda Ceriani. 2013. "The circadian system: Plasticity at many levels." *Neuroscience* 247:280-293. doi: 10.1016/j.neuroscience.2013.05.036.

Nair, Dinesh G. 2005. "About being BOLD." *Brain Research Reviews* 50 (2): 229-243. doi: 10.1016/j.brainreserv.2005.07.001.

Naneix, Benjamin. 2009. "The failure of the "localisationist project" in mental medicine in nineteenth century France and the emergence of the neurological clinic." *Poiesis und Praxis* 6 (1-2): 57-63. doi: 10.1007/s10202-008-0062-2.

Neubauer, Raymond L. 2014. "Prayer as an interpersonal relationship: A neuroimaging study." *Religion, Brain and Behavior* 4 (2): 92-103. doi: 10.1080/2153599X.2013.768288.

Newman, David V. 2011. "Chaos, emergence, and the mind-body problem." *Australasian Journal of Philosophy* 79 (2): 180-196. doi: 10.1080/713931202.

Newsome, William T. 2009. "Human freedom and "emergence"." In *Downward causation and the neurobiology of free will*, edited by Nancey Murphy, George F. R. Ellis and Timothy O'Connor, 55-62. Berlin: Springer.

Nichols, Thomas E., and Satoru Hayasaka. 2003. "Controlling the familywise error rate in functional neuroimaging: A comparative review." *Statistical Methods in Medical Research* 12 (5): 419-446. doi: 10.1191/0962280203sm341ra.

Nieder, Andreas, and Katharina Merten. 2007. "A labeled-line code for small and large numerosities in the monkey prefrontal cortex." *Journal of Neuroscience Methods* 27 (22): 5986-5993. doi: 10.1523/jneurosci.1056-07.2007.

Nieuwenhuis, Sander, Birte U. Forstmann, and Eric-Jan Wagenmakers. 2011. "Erroneous analyses of interactions in neuroscience: a problem of significance." *Nature Neuroscience* 14 (9): 1105-1107. doi: 10.1038/nn.2886.

Noë, Alva, and Evan Thompson. 2004. "Are there neural correlates of consciousness?" *Journal of Consciousness Studies* 11 (1): 3-28.

O'Connell, Garret, Janet De Wilde, Jane Haley, Kirsten Shuler, Burkhard Schafer, Peter Sandercock, and Joanna M. Wardlaw. 2011. "The brain, the science and the media - The legal, corporate, social and security implications of neuroimaging and the impact of media coverage." *EMBO Reports* 12 (7): 630-636. doi: 10.1038/embor.2011.115.

Page, Mike P.A. 2006. "What can't functional neuroimaging tell the cognitive psychologist?" *Cortex* 42 (3): 428-443. doi: 10.1016/S0010-9452(08)70375-7.

Pan, Hong, Jane Epstein, David A. Silbersweig, and Emily Stern. 2011. "New and emerging imaging techniques for mapping brain circuitry." *Brain Research Reviews* 67 (1-2): 226-251. doi: 10.1016/j.brainresrev.2011.02.004.

Paquette, Vincent, Johanne Lévesque, Boualem Mensour, Jean-Maxime Leroux, Gilles Beaudoin, Pierre Bourgouin, and Mario Beauregard. 2003. ""Change the mind and you change the brain": Effects of cognitive-behavioral therapy on the neural correlates of spider phobia." *NeuroImage* 18 (2): 401-409. doi: 10.1016/S1053-8119(02)00030-7.

Pardo, Michael S., and Dennis Patterson. 2010. "Philosophical foundations of law and neuroscience." *University of Illinois Law Review* 4:1211-1250.

Pashler, Harold, and Christine R. Harris. 2012. "Is the replicability crisis overblown? Three arguments examined." *Perspectives on Psychological Science* 7 (6): 531-536. doi: 10.1177/1745691612463401.

Perini, Laura. 2012. "Image interpretation: Bridging the gap from mechanically produced image to representation." *International Studies in the Philosophy of Science* 26 (2): 153-170. doi: 10.1080/02698595.2012.703478.

Petersen, Steven E., and Julie A. Fiez. 1993. "The processing of single words studied with positron emission tomography." *Annual Review of Neuroscience* 16:509-530. doi: 10.1146/annurev.ne.16.030193.002453.

Peterson, Bradley S. 2003. "Conceptual, methodological, and statistical challenges in brain imaging studies of developmentally based psychopathologies." *Development and Psychopathology* 15 (3): 811-832. doi: 10.1017/S0954579403000385.

Phan, K. Luan, Tor Wager, Stephan F. Taylor, and Israel Liberzon. 2002. "Functional neuroanatomy of emotion: A meta-analysis of emotion activation studies in PET and fMRI." *NeuroImage* 16 (2): 331-348. doi: 10.1006/nimg.2002.1087.

Poldrack, Russell A. 2006. "Can cognitive processes be inferred from neuroimaging data?" *Trends in Cognitive Sciences* 10 (2): 59-63. doi: 10.1016/j.tics.2005.12.004.

———. 2010. "Subtraction and beyond: The logic of experimental designs for neuroimaging." In *Foundational issues in human brain mapping*, edited by Stephen José Hanson and Martin Bunzl, 147-159. Cambridge, Mass: MIT Press.

Poldrack, Russell A., Paul C. Fletcher, Richard N. Henson, Keith J. Worsley, Matthew Brett, and Thomas E. Nichols. 2008. "Guidelines for reporting an fMRI study." *NeuroImage* 40 (2): 409-414. doi: 10.1016/j.neuroimage.2007.11.048.

Power, Jonathan D., Damien A. Fair, Bradley L. Schlaggar, and Steven E. Petersen. 2010. "The development of human functional brain networks." *Neuron* 67 (5): 735-748. doi: 10.1016/j.neuron.2010.08.017.

Price, Cathy J., and Karl J. Friston. 2005. "Functional ontologies for cognition: The systematic definition of structure and function." *Cognitive Neuropsychology* 22 (3-4): 262-275. doi: 10.1080/02643290442000095.

Quian Quiroga, Rodrigo, Gabriel Kreiman, Christof Koch, and Itzhak Fried. 2008. "Sparse but not "Grandmother-cell" coding in the medial temporal lobe." *Trends in Cognitive Sciences* 12 (3): 87-91. doi: 10.1016/j.tics.2007.12.003.

Quian Quiroga, Rodrigo, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. 2005. "Invariant visual representation by single neurons in the human brain." *Nature* 435 (7045): 1102-1107. doi: 10.1038/nature03687.

Qureshi, Irfan A., and Mark F. Mehler. 2013. "Towards a 'systems'-level understanding of the nervous system and its disorders." *Trends in Neuroscience* 36 (11): 674-684. doi: 10.1016/j.tins.2013.07.003.

Radecki, Guillaume, Romuald Nargeot, Ileana Ozana Jelescu, Denis Le Bihan, and Luisa Ciobanu. 2014. "Functional magnetic resonance microscopy at single-cell resolution in *Aplysia californica*." *Proceedings of the National Academy of Sciences of the United States of America* 111 (23): 8667-8672. doi: 10.1073/pnas.1403739111.

Raichle, Marcus E., Ann Mary MacLeod, Abraham Z. Snyder, William J. Powers, Debra A. Gusnard, and Gordon L. Shulman. 2001. "A default mode of brain function." *Proceedings of the National Academy of Sciences of the United States of America* 98 (2): 676-682. doi: 10.1073/pnas.98.2.676.

Raichle, Marcus E., and Mark A. Mintun. 2006. "Brain work and brain imaging." *Annual Review of Neuroscience* 29:449-476. doi: 10.1146/annurev.neuro.29.051605.112819.

Ramani, Donato. 2009. "The brain seduction: the public perception of neuroscience." *Journal of Science Communication* 8 (4): L01.

Ransohoff, David F., and Richard M. Ransohoff. 2001. "Sensationalism in the media: when scientists and journalists may be complicit collaborators." *Effective Clinical Practice* 4 (4): 185-188.

Rathkopf, Charles A. 2013. "Localization and intrinsic function." *Philosophy of Science* 80 (1): 1-21. doi: 10.1086/668878.

Reeves, Donald, Mark J. Mills, Stephen B. Billick, and Jonathan D. Brodie. 2003. "Limitations of brain imaging in forensic psychiatry." *Journal of the American Academy of Psychiatry and the Law* 31 (1): 89-96.

Regatte, Ravinder R. 2014. "Why buy an expensive ($7 Million) 7T MRI system for biomedical research?" *Journal of Magnetic Resonance Imaging* 40 (2): 280-282. doi: 10.1002/jmri.24444.

Reiner, Peter B. 2011. "The rise of neuroessentialism." In *The Oxford handbook of neuroethics*, edited by Judy Illes and Barbara J. Sahakian, 161-176. Oxford: Oxford University Press.

Robillard, Julie M., and Judy Illes. 2011. "Lost in translation: neuroscience and the public." *Nature Reviews Neuroscience* 12 (2). doi: 10.1038/nrn2773-c2.

Rose, Nikolas , and Joelle M. Abi-Rached. 2013. *Neuro: The new brain sciences and the management of the mind*. Princeton, NJ: Princeton University Press.

Rose, Steven. 1999. "Précis of *Lifelines: Biology, freedom, determinism*." *Behavioral and Brain Sciences* 22 (5): 871-885.

Roskies, Adina L. 2008. "Neuroimaging and inferential distance." *Neuroethics* 1 (1): 19-30. doi: 10.1007/s12152-007-9003-3.

———. 2010. "Neuroimaging and inferential distance: The perils of pictures." In *Foundational issues in human brain mapping*, edited by Stephen José Hanson and Martin Bunzl, 195-215. Cambridge, Mass.: MIT Press.

Rossier, Jean. 2009. "Wiring and plumbing in the brain." *Frontiers in Human Neuroscience* 3 (2). doi: 10.3389/neuro.09.002.2009.

Rueger, Alexander. 2000. "Physical emergence, diachronic and synchronic." *Synthese* 124 (3): 297-322. doi: 10.1023/A:1005249907425.

Sacchet, Matthew D., and Brian Knutson. 2013. "Spatial smoothing systematically biases the localization of reward-related brain activity." *NeuroImage* 66:270-277. doi: 10.1016/j.neuroimage.2012.10.056.

Sadaghiani, Sepideh, Guido Hesselmann, Karl J. Friston, and Andreas Kleinschmidt. 2010. "The relation of ongoing brain activity, evoked neural responses, and cognition." *Frontiers in Systems Neuroscience* 4 (20). doi: 10.3389/fnsys.2010.00020.

Saks, Michael J., N. J. Schweitzer, Eyal Aharoni, and Kent A. Kiehl. 2014. "The impact of neuroimages in the sentencing phase of capital trials." *Journal of Empirical Legal Studies* 11 (1): 105-131. doi: 10.1111/jels.12036.

Sawahata, Yasuhito, Kazuteru Komine, Toshiya Morita, and Nobuyuki Hiruma. 2013. "Decoding humor experiences from brain activity of people viewing comedy movies." *PLoS ONE* 8 (12): e81009. doi: 10.1371/journal.pone.0081009.

Saxe, Rebecca, Matthew Brett, and Nancy Kanwisher. 2010. "Divide and conquer: A defense of functional localizers." In *Foundational issues in human brain mapping*, edited by Stephen José Hanson and Martin Bunzl, 25-41. Cambridge, Mass.: MIT Press.

Schomerus, Georg, C. Schwahn, Anita Holzinger, P. W. Corrigan, H. J. Grabe, Mauro G. Carta, and Matthias C. Angermeyer. 2012. "Evolution of public attitudes about mental illness: A systematic review and meta-analysis." *Acta Psychiatrica Scandinavica* 125 (6): 440-452. doi: 10.1111/j.1600-0447.2012.01826.x.

Schreiber, Darren, Greg Fonzo, Alan N. Simmons, Christopher T. Dawes, Taru Flagan, James H. Fowler, and Martin P. Paulus. 2013. "Red brain, blue brain: Evaluative processes differ in democrats and republicans." *PLoS ONE* 8 (2): e52970. doi: 10.1371/journal.pone.0052970.

Schwartz, Lisa M, Steven Woloshin, Alice Andrews, and Therese A Stukel. 2012. "Influence of medical journal press releases on the quality of associated newspaper coverage: Retrospective cohort study." *British Medical Journal* 344 (d8164). doi: 10.1136/bmj.d8164.

Schweitzer, N. J., D Baker, and Evan Risko, F. 2013. "Fooled by the brain: Re-examining the influence of neuroimages." *Cognition* 129 (3): 501-511. doi: 10.1016/j.cognition.2013.08.009.

Seixas, Daniela, and Margarida Ayres Basto. 2008. "Ethics in fMRI studies: A review of the EMBASE and MEDLINE literature." *Clinical Neuroradiology* 18 (2): 79-87. doi: 10.1007/s00062-008-8009-5.

Shifferman, Eran M. 2011. *The evolution of quantity estimation in the animal kingdom*, Faculty of Humanities, School of Philosophy, Tel Aviv University, Tel Aviv.

Siero, Jeroen C. W., Dora Hermes, Hans Hoogduin, Peter R. Luijten, Natalia Petridou, and Nick F. Ramsey. 2013. "BOLD consistently matches electrophysiology in human sensorimotor cortex at increasing movement rates: a combined 7T fMRI and ECoG study on neurovascular coupling." *Journal of Cerebral Blood Flow & Metabolism* 33 (9): 1448-1456. doi: 10.1038/jcbfm.2013.97.

Silberstein, Michael. 2002. "Reduction, emergence and explanation." In *The Blackwell guide to the philosophy of science*, edited by Peter Machamer and Michael Silberstein, 80-107. Oxford: Blackwell.

Singh, Krish D. 2012. "Which "neural activity" do you mean? fMRI, MEG, oscillations and neurotransmitters." *NeuroImage* 62 (2): 1121-1130. doi: 10.1016/j.neuroimage.2012.01.028.

Sirotin, Yevgeniy B., and Aniruddha Das. 2009. "Anticipatory haemodynamic signals in sensory cortex not predicted by local neuronal activity." *Nature* 457 (7228): 475-479. doi: 10.1038/nature07664.

Smith, Alec, Terry Lohrenz, Justin King, P. Read Montague, and Colin F. Camerer. 2014. "Irrational exuberance and neural crash warning signals during endogenous experimental market bubbles." *Proceedings of the National Academy of Sciences of the United States of America* 111 (29): 10503-10508. doi: 10.1073/pnas.1318416111.

Smith, Philip L., and Roger Ratcliff. 2004. "Psychology and neurobiology of simple decisions." *Trends in Neurosciences* 27 (3): 161-168. doi: 10.1016/j.tins.2004.01.006.

Sporns, Olaf. 2014. "Contributions and challenges for network models in cognitive neuroscience." *Nature Neuroscience* 17 (5): 652-660. doi: 10.1038/nn.3690.

Sporns, Olaf, and Rolf Kötter. 2004. "Motifs in brain networks." *PLoS Biology* 2 (11): e369. doi: 10.1371/journal.pbio.0020369.

Sprevak, Mark. 2011. "Neural sufficiency, reductionism, and cognitive neuropsychiatry." *Philosophy, Psychiatry & Psychology* 18 (4): 339-344. doi: 10.1353/ppp.2011.0057.

Stelzer, Johannes, Gabriele Lohmann, Karsten Mueller, Tilo Buschmann, and Robert Turner. 2014. "Deficient approaches to Human neuroimaging." *Frontiers in Human Neuroscience* 8. doi: 10.3389/fnhum.2014.00462.

Sterne, Jonathan A. C., D. R. Cox, and George Davey Smith. 2001. "Sifting the evidence - what's wrong with significance tests?Another comment on the role of statistical methods." *British Journal of Medicine* 322 (7280): 226-231. doi: 10.1136/bmj.322.7280.226.

Suleski, Julie, and Motomu Ibaraki. 2010. "Scientists are talking, but mostly to each other: A quantitative analysis of research represented in mass media." *Public Understanding of Science* 19 (1): 115-125. doi: 10.1177/0963662508096776.

Tancredi, Laurence R., and Jonathan D. Brodie. 2007. "The brain and behavior: Limitations in the legal use of functional magnetic resonance imaging." *American Journal of Law and Medicine* 2-3 (271-294).

Thyreau, Benjamin, Yannick Schwartz, Bertrand Thirion, Vincent Frouin, Eva Loth, Sabine Vollstädt-Klein, Tomas Paus, Eric Artiges, Patricia J. Conrod, Gunter Schumann, Robert Whelan, and Jean-Baptiste Poline. 2012. "Very large fMRI study using the IMAGEN database: Sensitivity-specificity and population effect modeling in relation to the underlying anatomy." *NeuroImage* 61 (1): 295-303. doi: 10.1016/j.neuroimage.2012.02.083.

Tomassy, Giulio Srubek, Daniel R. Berger, Hsu-Hsin Chen, Narayanan Kasthuri, Kenneth J. Hayworth, Alessandro Vercelli, H. Sebastian Seung, Jeff W. Lichtman, and Paola Arlotta. 2014. "Distinct profiles of myelin distribution along single axons of pyramidal neurons in the neocortex." *Science* 344 (6181): 319-324. doi: 10.1126/science.1249766.

Uttal, William R. 2002. "Précis of *The new phrenology*: The limits of localizing cognitive processes in the brain." *Brain and Mind* 3 (2): 221-228. doi: 10.1023/A:1019972122144.

———. 2013. *Reliability in cognitive neuroscience: A meta-meta-analysis*. Cambridge, MA: MIT Press.

van Orden, Guy C., and Kenneth R. Paap. 1997. "Functional neuroimages fail to discover pieces of mind in parts of the brain." *Philosophy of Science* 64 (4): S85-S94.

Vidal, Fernando. 2009. "Brainhood, anthropological figure of modernity." *History of the Human Sciences* 22 (1): 5-36. doi: 10.1177/0952695108099133.

Vul, Edward, Christine Harris, Piotr Winkielman, and Harold Pashler. 2009. "Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition." *Perspectives on Psychological Science* 4 (3): 274-290. doi: 10.1111/j.1745-6924.2009.01125.x.

Vul, Edward, and Nancy Kanwisher. 2010. "Begging the question: The nonindependence error in fMRI data analysis." In *Foundational issues in human brain mapping*, edited by Stephen José Hanson and Martin Bunzl, 71-91. Cambridge, Mass.: MIT Press.

Vul, Edward, and Hal Pashler. 2012. "Voodoo and circularity errors." *NeuroImage* 62 (2): 945-948. doi: 10.1016/j.neuroimage.2012.01.027.

Wacholder, Sholom, Stephen Chanock, Montserrat Garcia-Closas, Laure El ghormli, and Nathaniel Rothman. 2004. "Assessing the probability that a positive report is false: An approach for molecular epidemiology studies." *Journal of the National Cancer Institute* 96 (6): 434-442. doi: 10.1093/jnci/djh075.

Wager, Tor D., Martin A. Lindquist, Thomas E. Nichols, Hedy Kober, and Jared X. Van Snellenberg. 2009. "Evaluating the consistency and specificity of neuroimaging data using meta-analysis." *NeuroImage* 45 (1, Supplement 1): S210-S221. doi: 10.1016/j.neuroimage.2008.10.061.

Wan, Ming Wai, Darragh Downey, Hilary Strachan, Rebecca Elliott, Steve R. Williams, and Kathryn M. Abel. 2014. "The neural basis of maternal bonding." *PLoS ONE* 9 (3): e88436. doi: 10.1371/journal.pone.0088436.

Wang, Xiao-Jing. 2010. "Neurophysiological and computational principles of cortical rhythms in cognition." *Physiological Reviews* 90 (3): 1195-1268. doi: 10.1152/physrev.00035.2008.

Wardlaw, Joanna M., Garret O'Connell, Kirsten Shuler, Janet DeWilde, Jane Haley, Oliver Escobar, Shaun Murray, Robert Rae, Donald Jarvie, Peter Sandercock, and Burkhard Schafer. 2011. ""Can it read my mind?" - What do the public and experts think of the current (mis)uses of neuroimaging?" *PLoS ONE* 6 (10): e25829. doi: 10.1371/journal.pone.0025829.

Whitelely, Louise. 2012. "Resisting the revelatory scanner: Critical engagements with fMRI in popular media." *BioSocieties* 7 (3): 245-272. doi: 10.1057/biosoc.2012.21.

Whittington, Miles A., Mark O. Cunningham, Fiona E. N. LeBeau, Claudia Racca, and Roger D. Traub. 2011. "Multiple origins of the cortical gamma rhythm." *Developmental Neurobiology* 71 (1): 92-106. doi: 10.1002/dneu.20814.

Woloshin, Steven, and Lisa M. Schwartz. 2002. "Press releases: Translating research into news." *Journal of the American Medial Association* 287 (21): 2856-2858. doi: 10.1001/jama.287.21.2856.

Wolpe, Paul Root. 2006. "Reasons scientists avoid thinking about ethics." *Cell* 125 (6): 1023-1025. doi: 10.1016/j.cell.2006.06.001.

Wouters, Arno. 2005. "The function debate in philosophy." *Acta Biotheoretica* 53 (2): 123-151. doi: 10.1007/s10441-005-5353-6.

Wynne, Brian. 2001. "Creating public alienation: Expert cultures of risk and ethics on GMOs." *Science as Culture* 10 (4): 445-481. doi: 10.1080/09505430120093586.

Yang, Zhi, Zirui Huang, Javier Gonzalez-Castillo, Rui Dai, Georg Northoff, and Peter Bandettini. 2014. "Using fMRI to decode true thoughts independent of intention to conceal." *NeuroImage* 99 (0): 80-92. doi: 10.1016/j.neuroimage.2014.05.034.

Yarkoni, Tal. 2009. "Big correlations in little studies: Inflated fMRI correlations reflect low statistical power - Commentary on Vul et al. (2009)." *Perspectives on Psychological Science* 4 (3): 294-298. doi: 10.1111/j.1745-6924.2009.01127.x.

Yoder, Keith J., and Jean Decety. 2014. "The good, the bad, and the just: Justice sensitivity predicts neural response during moral evaluation of actions performed by others." *The Journal of Neuroscience* 34 (12): 4161-4166. doi: 10.1523/jneurosci.4648-13.2014.

Yong, Ed. 2012. "Replication studies: Bad copy." *Nature* 485 (7398): 298-300. doi: 10.1038/485298a.

Yuste, Rafael. 2013. "Electrical compartmentalization in dendritic spines." *Annual Review of Neuroscience* 36:429-449. doi: 10.1146/annurev-neuro-062111-150455.

Zandbelt, Bram B., Thomas E. Gladwin, Mathijs Raemaekers, Mariët van Buuren, Sebastiaan F. Neggers, René S. Kahn, Nick F. Ramsey, and Matthijs Vink. 2008. "Within-subject variation in BOLD-fMRI signal changes across repeated measurements: Quantification and implications for sample size." *NeuroImage* 42 (1): 196-206. doi: 10.1016/j.neuroimage.2008.04.183.

Zeki, Semir. 2015. "A massively asynchronous, parallel brain." *Philosophical Transactions of the Royal Society B - Biological Sciences* 370 (1668). doi: 10.1098/rstb.2014.0174.

Zuckerman, Diana. 2003. "Hype in health reporting: "checkbook science" buys distortion of medical news." *International Journal of Health Services* 33 (2): 383-389. doi: 10.2190/pmm9-dput-hn3y-lmjq.

Journal of Cognition and Neuroethics

The Theory-Theory of Moral Concepts

John Jung Park
Christopher Newport University

Biography
My interests lie in moral psychology, metaethics, and philosophy of mind. I am particularly interested in the structure of our moral concepts and what implications this has for metaethics. I am also interested in the hard problem of consciousness.

Acknowledgements
I would like to thank David Wong, Karen Neander, Daniel Weiskopf, and Wayne Norman for their input on this paper.

Publication Details
Journal of Cognition and Neuroethics (ISSN: 2166-5087). September, 2015. Volume 3, Issue 1.

Citation
Park, John Jung. 2015. "The Theory-Theory of Moral Concepts." *Journal of Cognition and Neuroethics* 3 (1): 117–138.

The Theory-Theory of Moral Concepts

John Jung Park

Abstract

There are many views about the structure of concepts, a plausible one of which is the theory-theory. Though this view is plausible for concrete concepts, it is unclear that it would work for abstract concepts, and then for moral concepts. The goal of this paper is to provide a plausible theory-theory account for moral concepts and show that it is supported by results in the moral psychology literature. Such studies in moral psychology do not explicitly contend for the theory-theory of moral concepts, but I demonstrate that they actually do provide evidence for the use of theory knowledge at times in moral categorization and decision-making. In philosophy of cognitive science, I newly show that there is evidence that the theory-theory does apply to some moral concepts.

Keywords

Moral Concepts, Concepts, Mental Representations, Theory-Theory, Prototype Theory, Moral Cognition, Empirical Moral Psychology, Cognitive Science

Introduction

The theory view for concrete concepts claims that concrete concepts are mental representations of hidden essences, causal laws, functions, explanatory relations, and/or general background knowledge (Carey 1985 and 2009; Murphy & Medin 1985; Keil 1989; Gopnik & Meltzoff 1997).[1] As my principle aims in this paper are to freshly explicate the theory view for moral concepts and provide evidence for it in light of moral categorization, I perceive my contribution in this case to be primarily in the concepts literature, where theorists in part address what constitutes concepts (Rosch & Mervis 1975; Prinz 2002; Machery 2009; Weiskopf 2009), rather than in the causal moral judgment literature in moral psychology, where moral psychologists examine what mental states influence the making of moral judgments (Greene et al. 2009; 2013; Mendez 2005; Cushman 2008; Young & Saxe 2008; Haidt 2012). The first reason for this is that the causal judgment literature rarely explicitly discusses the structure of moral concepts. Second, some of the main aims of this paper are to elaborate upon the theory view in the concepts literature and show how such a view applies at times to the

1. Mental representations are mental states that refer to or purport to represent things in the world. For example, my concept DOG refers to the category *dog*.

moral concepts domain. Hence, I see my contribution that establishes the viability of the theory-theory for moral concepts as being primarily in the concepts field. To note, I do not claim that moral concepts only have theory structure. I perfectly leave open the possibility that moral concepts can store many other different kinds of knowledge that can be used individually or conjointly in cognition. However, due to obvious space concerns, I will focus exclusively on the issue of whether some moral concepts may have theory structure.

Psychological concepts are generally understood as being the constituents of thought or as Locke states, they are the "materials of reason and knowledge." They are the basic units of the human understanding. For example, my judgment RAPE IS WRONG is made up of three individual concepts: RAPE, IS, and WRONG.[2] Also, concepts are understood as being mental representations or bodies of knowledge[3] that are stored in long term memory and are functionally used in most of the higher cognitive competences, where the relevant competences are such things as categorization, induction, deduction, concept combination, and planning (Machery 2009; Weiskopf 2009). While there are alternate notions of a concept,[4] in this paper we will understand concepts in the widely understood sense in cognitive science and philosophical psychology. Here, concepts are the constituents of thought and mental representations used in most of the higher competences.

Concepts are theoretical constructs in psychology that play a pivotal role in explaining higher acts of cognition. Moral concepts and their structures are in significant part responsible for how we perform competences in moral cognition such as moral

2. I follow normal convention in the concepts literature and capitalize all concepts.

3. Throughout this paper, 'knowledge' will stand for an information-carrying mental state rather than the traditional philosophical understanding of true justified belief. In this respect, I follow standard convention in the concepts literature.

4. Fodor's informational atomism theory of concepts is meant to provide a theory of content for concepts and not necessarily provide a view of how concepts partake in the higher competences (Fodor 1998). Providing a theory of content and a theory of how concepts play a functional role in higher cognition can be seen as being in principle two different projects, although one may pursue both projects. Also, the notion of a concept used here differs from the idea of a platonic concept that is an abstract object rather than a mental representation. While platonic concepts are focused on the metaphysically correct features of a category, the interest in this paper will be on epistemic mental representation concepts in the minds of human beings that may change over time and may be incorrect. Finally, some psychological concepts can be personal states and others can be subpersonal states. Insofar as personal states have the function of being intentionally used and subpersonal states do not have this function, psychological concepts can be either of the two kinds of states.

categorization, decision-making, planning, analogical reasoning, and induction. However, very few concept theorists have worked on the nature of abstract moral concepts. No experiments have been run explicitly on the theory-theory for moral concepts, and it is not even clear what theory knowledge is in the moral concepts domain. Despite the central importance of moral concepts for moral cognition, surprisingly, only a few moral psychologists explicitly have worked on this project.[5] The causal judgment literature in moral psychology that examines what mental states influence moral decision-making rarely discusses the concepts literature, and the causal judgment literature generally does not explicitly examine the structure of moral concepts. The causal judgment literature in moral psychology generally does not discuss the main player in moral cognition and moral decision-making in the psychological domain; namely, moral concepts. However, as we shall later see, since concepts and their psychological structures are defined as influencing decision-making and the causal judgment literature examines what mental states influence moral decision-making, I will use the causal judgment literature in moral psychology to draw certain conclusions about the structure of some moral concepts.

Although when asked, concept theorists likely will not deny the possibility that moral concepts might have theory structure, it has yet to be stated and proven in the literature that moral concepts indeed do have theory structure. Several studies have demonstrated that some abstract concepts may have different structures than those commonly found in concrete concepts (Hampton 1981; Barsalou et al. 2005; Wiemer-Hastings et al. 2005). For instance, James Hampton, ran tests on eight different abstract concepts, such as BELIEF, SCIENCE, and CRIME, along with tests on some concrete concepts in order to determine whether the abstract concepts had prototype structure similar to the successful results of finding prototype structure in concrete concepts. Although we will elaborate on this theory later, the prototype view claims that concepts are constituted by prototypes or mental representations of the statistically frequent features of members of a class (Rosch and Mervis 1975). In Hampton's study, the results were a mixed bag where some abstract concepts did show prototype structure, but others did not. For example, SCIENCE and CRIME showed prototype structure while abstract concepts such as A BELIEF and AN INSTINCT, that may intuitively be thought to have prototype structure, as a matter of fact do not have such structure. Thus, the upshot from Hampton's and others' experiments is that we cannot safely presuppose that abstract concepts will have the same theoretical concept structure and cognitive processing as those for concrete

5. Some exceptions are Jesse Prinz (2008), Stephen Stich (1993), Mark Johnson (1993), Paul Churchland (1989), and David Wong (2006).

concepts. As a matter of caution, we cannot draw conclusions about moral concepts solely based on the findings of concrete concepts. Therefore, further work is required in order to ascertain the structure of moral concepts, such as whether some of them have theory structure. I will put forth this further work in order to demonstrate the viability of the theory-theory for some moral concepts.

The Theory-Theory for Concrete Concepts

We will now elaborate on the theory-theory for concrete concepts in order to construct and get a proper sense of how the theory-theory may look like for moral concepts. The theory-theory of concepts is a view that emerged out of psychology in the 1980s, although in philosophy it has its roots in the likes of Locke (1689) and Quine (1977). The theory-theory of concepts states that concepts are themselves theories or mini-theories. Theories or mini-theories are certain mental representations. More specifically, they are scientific, hidden essence, causal law-like, functional, explanatory, and general or generic background knowledge about the extension of a concept and can explain such things as categorization in concrete concepts. For example, Edouard Machery, who is a proponent of this view *inter alia*, writes:

> Psychologists assume that laws, causal propositions, functional propositions…, and generic propositions…explain why things happen. Thus, a theoretical concept is supposed to store some nomological, causal, functional, and/or generic knowledge about the members of its extension. (2009, 101)

A description of these various kinds of theory mental representations will be discussed below in turn.

It will help to understand the theory view, especially for those who are not familiar with the concepts literature, by contrasting it with the prototype theory of concepts. As we shall later see, a major shortcoming of the prototype view is its inability to account for the additional knowledge that has been found to be stored in many concepts that accounts for some cases of classification. Recall that the prototype view claims that concepts are constituted by or just *are* prototypes or mental representations that refer to the statistically frequent features of members of a class. By the term 'constituted,' I am referring to the 'is' of identity. For example, Jesse Prinz, who is a proponent of this view *inter alia*, states that, "[M]any categories are associated with small sets of "typical" features. Typical features are ones that are diagnostic, statistically frequent, or salient. Unlike defining features, they are often contingent for category membership" (2002, 52).

Inspiration for the prototype view explicitly comes from Wittgenstein's notion of family resemblance, where members of a category may have one to several features or characteristics in common with each other, but zero or very few characteristics are common to all category members.[6] Prototype features are considered statistically frequent in that it is highly probable that a member of one's category will have them. Prototypes represent the superficial appearances of an object. Moreover, prototypes do not represent features that are necessary and sufficient conditions for determining membership. For example, one's prototype of DOG may be HAS HAIR, HAS FOUR LEGS, BARKS, PLAYS FETCH, and WAGS ITS TAIL. The features that are represented may be arrived upon based on all of one's previous experiences with particular dogs. Such prototypes can influence categorization decisions of whether something is or is not a dog. Furthermore, a dog may still be considered a dog even though it is hairless or even if it only has three legs.

On prototype views, features may be weighted more heavily than others. For example, *barks* and *plays fetch* may be weighed more heavily than *has hair*. When calculating how similar a potential or target member may be to a category, a prototype theory may take into account the number of features the instance may share with a category, or the instance's satisfaction of heavily weighted features, or both. Prototype theory is considered a similarity-based view because when an object or act is similar enough to the representation of summary or general features and passes a calculated similarity threshold, then the object falls within the class. Returning to the example, since my pet animal satisfies the importantly heavily weighted features of *barks* and *plays fetch*, along with several of the other features, it passes the similarity threshold and is categorized as a dog. When a token passes the similarity threshold of two or more categories, it is generally categorized in the class towards which it has the highest similarity score. At the same time, passing the similarity threshold for multiple categories can also beneficially explain the phenomena of ambiguous cases of membership, where some individuals may categorize an item as a member of two different classes.

The theory-theory attempts to address this issue of the superficial nature of the features represented by prototypes by claiming that background knowledge of the world rather than prototypes that are about superficial properties can play a role in higher acts of cognition. In other words, theories or mental representations of things like hidden essences, causal laws, and functions can determine how we make categorization

6. There are various versions of the prototype theory and disagreements between different proponents of such views. However, I detail here the prototype theory as it generally may be understood.

decisions. Use of theory knowledge as well as other conceptual knowledge, such as prototypes, in cognition may be effortful, automatic, conscious, or subconscious.

In an experiment for the theory-theory, Frank Keil ran a study where participants were asked whether the animal in a given scenario is a horse or a cow (1989, 162). In the situation, there is an animal that is called a 'horse,' makes horse sounds, looks like a horse, is strapped with a saddle so people can ride on it, and eats oats and hay. The animal has all the superficial prototypical features of a horse. However, scientists run blood tests and x-rays on it, and they discover that its insides are actually the insides of a cow. In this experiment, Keil found that older children and adults perceived the scientists' discoveries as relevant for determining natural kind membership. These subjects relied not on superficial similarities but on folk biological theories of hidden essences to decide that the animal was really a cow despite its superficial horse appearances.

As an example of the importance and use of folk causal law knowledge in cognition, *being curved* is an equally typical prototype feature in bananas and boomerangs. However, subjects give more weight to this attribute in boomerangs rather than bananas because it is falsely believed that curvature is causally related to the boomerang's property of *if thrown, it will return back to the thrower* (Medin & Shobin 1988). Due to this relationship between the two features, it is thought that *being curved* is more required for a boomerang rather than a banana. Theories may provide other causal explanatory relations between superficial features of an object. As an example that is an oversimplification for the point of illustration, my FISH concept may be constituted by the prototypes: HAS FINS, HAS A TAIL, and SWIMS (Murphy & Medin 1985). Theoretical knowledge of fish provides the explanation of the relation between fish attributes since in order to properly swim, a fish needs fins and a tail. Furthermore, participants believe that the hidden essence of a natural kind is generally causally responsible for the superficial features of the kind. For instance, many believe that the hidden essence of human beings is responsible for why we have the typical observable properties that we do. Here, most theory-theorists usually do not necessarily deny that one may have in mind superficial features when representing a class (Medin & Shobin 1988; Murphy & Medin 1985; Carey 2009), but they do emphasize the importance of such things as folk causal law knowledge or theories in providing the underlying explanation to such features as well as in deciding what weight such features may possess.

Theory-theorists also hold that there are domain differences for types of knowledge where different ontological domains contain different types of central beliefs. For example, while natural kinds are believed to have hidden essences, the analogue for artifact kinds generally is intended function. For example, Lin and Murphy ran an

experiment where they first described and showed pictures of certain artifacts from foreign countries (1997). One such item is a *tuk*. A tuk is a hunting tool that is a stick with a special handle on one end that protects the wielder's hand from animal bites. On the other end of the stick is a noose that goes around the head of the animal. The function of the tuk is to be able to control an animal by placing the noose over its head. After informing participants about what a tuk is and the function it performs, the experimenters showed participants a picture of what looks like a tuk minus only the special handle. When asked to categorize the item, participants categorized it as a tuk. When shown a picture of what looks like a tuk minus only the noose, subjects did not categorize it as a tuk. This suggests that functional knowledge plays a role in categorization, where participants did not categorize the latter item as a tuk because it could not perform the tuk's function.

Moreover, theory knowledge may not only contain knowledge of hidden essences, theoretical entities, and causal laws, but they may also contain general background knowledge. Theory knowledge need not be restricted to those kinds of knowledge that are of properties that are related to the structure of scientific theories, properties such as causal laws and essences. For example, Murphy and Medin claim that most people think the feature of *flammable* is a quality of wood rather than paper money even though both wood and paper money are flammable (1985). The reason behind this is that we have general background knowledge about the world concerning human activity where wood is used for burning fires and paper money is mostly used for economic purposes in which its flammability plays no role. On their view, this knowledge still counts as a theory that influences feature attribution for classes even though it may not be about a causal law or hidden essence.[7] For, such general background knowledge about human activity in principle still does provide an *explanation* of why we may attribute certain features to certain classes.

The Theory-Theory for Moral Concepts

In this section, we will strike new ground in detailing how the theory view will look for moral concepts. In order to properly discuss how we may make the appropriate changes to the theory-theory in order to account for the moral domain and moral concepts, it will help to discuss the prototype theory of moral concepts. Just as we used

7. In this respect, their understanding of the theory-theory differs from the likes of Gopnik and Meltzoff who draw a much tighter connection between theory conceptual structure and properties that are related to the structure of scientific theories.

the prototype theory above to help illustrate how the theory view posits background knowledge conceptual structures to concrete concepts that underlie superficial prototypes, we will likewise use the prototype theory of moral concepts in order to help illustrate the import and nature of the theory view for moral concepts.

If prototype theory is viable for moral concepts, then the prototype for the moral concept RIGHT ACTION for some individual may be BEING GENEROUS TO OTHERS, HELPING THE HOMELESS IS THE RIGHT THING TO DO BECAUSE IT BENEFITS THOSE IN NEED, PREVENT HARM, DOES NOT BREAK LAWS, and EXHIBITS FRIENDLINESS (Walker & Pitts 1998; Walker & Hennig 2004; Park 2013). Such representations, when understood as features of members of a class, are not necessary and sufficient conditions. As we can see, prototypes may be about such things as general features of moral situations, virtues, reasons for action, and basic moral principles or rules. For example, when a person points out to another an instance where a stranger is helping homeless people that such is a case of moral rightness, the mentally represented reason or justification for action that HELPING THE HOMELESS SO LONG AS ONE IS NOT IN POVERTY ONESELF IS THE CORRECT THING TO DO BECAUSE IT BENEFITS THOSE IN NEED may now be a candidate to be a constituent component of this listener's prototype of RIGHT ACTION based on further particular experiences. The abstraction of such features may be based on personal experiences and moral education. Now, for the individual in question, the virtue friendliness may carry less weight for this individual as compared to heavily weighted features such as the principles *prevent harm* and *does not break laws*.

Concerning the theory-theory as applied to ethical concepts, ethical concepts may themselves be theories that have as components such things as knowledge about master moral principles from which other moral principles generally may be thought to be inferred and explained. For example, the concept RIGHT ACTION may be constituted by normative theoretical information akin to divine command theory. The theory ACT IN ACCORDANCE WITH THE PRINCIPLES MANDATED BY GOD can be a component of RIGHT ACTION.

As previously stated, prototypes are about features that may be such things as moral principles, reasons for action, and virtues. However, the theory-theory components are more about master moral principles from which other moral principles, reasons, and virtues generally may be thought to be inferred and explained. For example, from divine command theory, where one must obey those laws mandated by God, one may arrive upon principles such as *do not lie* and *do not steal*. One may adhere to these moral principles based on one's background belief in the ethical theory of divine command

theory. Moreover, for this person, divine command theory explains why we must not lie and steal. On the other hand, one may have a virtue ethics master moral principle in mind such as EXEMPLIFY THOSE VIRTUES THAT THE VIRTUOUS PERSON HAS from which one may infer the proper virtues, such as KINDNESS, HONESTY, GENEROSITY, and PATIENCE. Moreover, this theory knowledge or master principle explains why the particular individual adheres to the group of virtues that she does. Also, a person may have in mind the universalizability principle as theory knowledge to infer maxims. The mentally represented deontological master principle OBEY THAT MAXIM THAT ONE CAN WILL TO BE A UNIVERSAL LAW can be used by a particular person to infer and explain the set of maxims this individual holds, such as DO NOT LIE and DO NOT STEAL.

Also, the act utilitarian Greatest Happiness Principle, where one must perform that action that leads to the greatest happiness for the greatest number, initially appears to be only about generating verdicts on acts rather than deriving other principles. However, act utilitarianism may not produce formal pithy rules or principles but it still relies on reasons for action to arrive upon its act-based conclusions. As a utilitarian, in self-defense one may believe that one ought to take the life of a murderous assailant. However, this may be based on the specified reason that in regards to overall happiness, killing others is wrong, but in acts of self-defense, killing others is justifiable. Such a reason may be thought to stem from and be explained by utilitarianism in that it is under the eye of the Greatest Happiness Principle that such a reason or justification is formed and used to determine what act one should follow. Thus, act utilitarianism may also be thought to be a theory or master principle from which reasons or considerations that count in favor of something are based.

While the theory-theory is a distinct view from the prototype theory based on the specific kind of knowledge it posits concepts as containing, as we can see, there is an intimate link between the theory-theory and the prototype theory for moral concepts. Ethical theory knowledge provides an underlying explanation for why one holds the moral prototypes that one does. This is just like the intimate link between both concept theories in the concrete concept realm, where as shown in our previous example, general explanatory background knowledge is the reason why the folk attribute the prototypical property of *flammable* to wood but not paper money. Also, recall that Murphy and Medin showed how theory knowledge need not be restricted to those kinds of knowledge that are of properties related to the structure of scientific theories. Hence, theory knowledge for moral concepts also need not be about properties related to the structure of scientific theories.

It is knowledge about these master principles that belong to the theory-theory of moral concepts, while the moral principles, reasons, and virtues that may be thought to be inferable from master principles belong more in the domain of the prototype view. The reason why this is the case is that such master principles are those that ostensibly underlie the inferential principles, virtues, and reasons for action at a deeper level of theoretical abstraction just as, for example, folk biological theories of hidden essences underlie the superficial features of a biological natural kind. Second, and a related point, is that just as biological theories may be about *explanatory relations* between superficial features, master moral principle knowledge may provide an explanatory link between the inferential principles, virtues, or reasons. In this respect, inferential principles, reasons, and virtues may be thought of as being superficial. Meanwhile, master principles are more theoretical and lie at the deepest explanatory level. For example, *do not kill* and other rules such as *do not lie* may be ultimately explained and unified by divine command theory for a particular person. Represented master moral principles are theory knowledge precisely because they provide an underlying explanation for why one holds the inferential principles, virtues, or reasons for action that one does. This is just like how some theory knowledge of a class in the concrete concepts domain provides an underlying explanation of one's prototypes of that class.

What I have provided thus far for an explanation of the structure of the theory-theory for moral concepts may not exactly parallel the structural components of the theory-theory for concrete concepts given the differences between the abstract moral and concrete concept domains. However, in allying some theories with knowledge of master moral principles, we can see that several important similarities as just discussed exist between the given moral and concrete theory structures to warrant drawing the distinction between moral concept prototype and theory structural components in this manner.

Also, causal moral law knowledge about an agent's intentions for action may be theory components of moral concepts in that they may take part in the higher cognitive competences related to ethical matters. Just as natural kind concepts may contain causal law knowledge, moral concepts may also be constituted by causal moral law knowledge. For instance, in moral cognition, one may have a representation of a causal moral law or principle such as IF AN AGENT'S CAUSAL MOTIVATIONS ARE WRONG, THEN THE AGENT'S ACTION USUALLY IS WRONG. This may constitute one's WRONG concept. It may also underlie and be responsible for the prototype constituent of one's WRONG concept, LACKS MORAL WORTH, to carry significant weight and importance. If one does not perform an action with the proper motivations and intentions, then the act

lacks moral worth, and this may lead to the greater chance that the act is classified as a wrong act. Given the above causal moral law knowledge, LACKS MORAL WORTH may be held to contain significant weight and importance in one's WRONG concept. This is somewhat analogous to the theory-theory for concrete concepts in which one's BOOMERANG concept may be constituted by theory knowledge of the causal law-like principle IF THROWN, IT WILL RETURN BACK TO THE THROWER. Recall that this theory knowledge likewise underlies and impacts the weight carried by the prototype constituent of one's BOOMERANG concept, BEING CURVED. BEING CURVED is a more heavily weighted prototype component for BOOMERANG rather than BANANA because the causal principle knowledge contained in participants' BOOMERANG concept confers such additional weight.

In summary of this section, I have developed the theory-theory for moral concepts. Theory knowledge for moral concepts includes such things as ethical theory knowledge of master principles and causal moral law knowledge. Just as the use of hidden essence knowledge in the categorization of natural kinds provides evidence for the theory view in respect to such categorization, the use and influence of ethical theory and causal moral law knowledge in moral cognition when making moral categorization decisions of what is morally right and wrong will provide evidence for the theory view in respect to such tasks. Conceptual structures that constitute a concept, such as prototypes and theories, are posited as playing a functional or causal role in higher acts of cognition, such as in categorization and decision-making. In the subsequent section, we will first examine studies showing the use of ethical theory knowledge at times in moral categorization. Next, we will discuss several experiments that demonstrate the use of causal moral law knowledge at times in moral categorization. These various sets of studies to be examined each independently demonstrate the viability of the theory-theory for moral concepts.

As is becoming widely accepted in the concrete concepts literature (Prinz 2002; Murphy 2004; Machery 2009; Weiskopf 2009), I perfectly allow for the possibility that other kinds of knowledge other than theory knowledge can be used at times in cognition. For example, it may be the case that emotions are so used, which will lead to a view in which moral concepts are at least in part constituted by emotions. Prototype and other kinds of knowledge may also in part constitute our moral concepts, where different kinds of concept constituents can be used together or separately in moral cognition in different contexts. For instance, in one case of decision-making I may rely only on theory knowledge, and in another situation I may use only prototypes and emotions.

We will now examine several studies demonstrating that theory knowledge is also at times used in moral categorization. Although there may be other kinds of knowledge

used in moral cognition, due to space concerns, the focus here is precisely on the explication of the theory view in the moral concepts domain and providing evidence that theory knowledge is actually used at times in moral categorization.

Evidence for the Theory-Theory of Moral Concepts

Now that we have clarified what kind of knowledge is theory knowledge for moral concepts, we will examine well-known studies that are not explicitly designed by their authors to draw moral concept constitution conclusions, but they can indeed be used as evidence for the use of theory knowledge in moral categorization of what is morally right and wrong. As previously stated, there are no explicit studies on the theory-theory for moral concepts in the concepts literature or in moral psychology. However, there are numerous experiments in moral psychology that provide evidence for the theory-theory of moral concepts even though the relevant literature does not at all discuss how such studies can be used to draw structural conclusions on moral concepts. The relevant connection has not been made in the relevant studies. In other words, there are several studies showing that at times, theory knowledge is used in moral categorization. However, in the literature, the connection is not drawn that since concepts are functionally defined, such studies provide evidence for the theory-theory of moral concepts. Another reason why this connection has not been made is that no one to this point has stated in the literature what theory knowledge even is for moral concepts. As we shall see, the use of theory knowledge in moral cognition is just like how in Keil's previously discussed study, participants used theory knowledge of hidden essences in order to make their categorization judgment that the given animal is really a cow instead of a horse.

Joshua Greene, et al. ran a cognitive load study where subjects were filling out moral questionnaires on a computer (2008). They were presented with "high conflict" moral dilemmas in which subjects were asked whether it is appropriate to harm another individual in order to save several lives. One example of a high conflict dilemma that was used is the crying baby case:

> Enemy soldiers have taken over your village. They have orders to kill all remaining civilians. You and some of your townspeople have sought refuge in the cellar of a large house. Outside you hear the voices of soldiers who have come to search the house for valuables.

> Your baby begins to cry loudly. You cover his mouth to block the sound. If you remove your hand from his mouth his crying will summon the attention of the soldiers who will kill you, your child, and the others hiding out in the cellar. To save yourself and the others you must smother your child to death.
>
> Is it appropriate for you to smother your child in order to save yourself and the other townspeople? (Greene et al. 2008, 1147–1148)

While answering such questions on moral dilemmas, numbers continually stream across the bottom of the screen and participants have to press a button when they see the number five. The result is that subjects selectively had a longer reaction time when making utilitarian judgments under the cognitive load as opposed to having no cognitive load, but there was no increase in reaction time for non-utilitarian judgments under cognitive load. This study provides causal rather than correlational support that utilitarian theory knowledge influences some moral judgments for some people because the longer reaction time suggests that the cognitive load of having to press a button when seeing the number five interferes with a controlled cognitive process, such as some kind of cost-benefit analysis reasoning process, whereas the cognitive load should have no effect on a fast automatic process.[8] This conclusion is further buttressed by neuroimaging

8. Some of the provided scenarios to participants use the famed trolley problems. While the likes of Greene and Haidt interpret the lever and footbridge cases as that between utilitarianism and deontology (Haidt 2012; Greene 2013), others such as Mikhail interpret both cases as involving the use of the Doctrine of Double Effect (Mikhail 2007). However, as is well documented, Greene and Haidt have pointed out that the Doctrine of Double Effect interpretation is false given results from a loop variant scenario in which participants will use a person as a means to save many lives (Haidt 2012; Greene 2013). In other words, subjects do not consistently abide by the Doctrine, and the opposite predictions from the Doctrine interpretation are borne out. Hence, I interpret studies reliant on the trolley problems to pit utilitarianism knowledge versus deontological knowledge. To note, even if the Doctrine was at work, its use in cognition would still provide evidence for the theory-theory; a concept constitution conclusion that is not foreseen nor anticipated in the writings of Mikhail. Mikhail, like other moral psychologists, fails to draw the connection in his works.

   Guy Kahane has argued in the trolley problems that the supposed evidence for the use of utilitarianism knowledge is really support for the use of deontological theory knowledge (2012). Yet, one major problem with Kahane's thesis is that 44 independently-run cross-cultural studies have demonstrated that there is extremely little correlation between moral judgments and deontological thinking, but there is a robust correlation between moral judgments and utilitarian thinking (for a summary of the studies, see Snarey 1985). While this robust correlation provides even further support for the particular theory-theory conclusion that utilitarian knowledge is used in moral cognition, without a general correlation between

studies that show a strong correlation between the making of utilitarian judgments and activation in the dorsolateral prefrontal cortex (Greene et al. 2004). This brain region is known for such things as complex planning, deductive/inductive reasoning, and long-term economic decision-making. The above provides evidence for the viability of the theory-theory for moral concepts.

Mendez and company have shown that frontotemporal dementia patients who have intact reasoning capacities but who have blunted affect or severely diminished emotions, tend to make the same utilitarian moral judgments as compared to normal subjects on the same moral scenarios (2005). Moreover, Koenigs, Young, and company (2007) as well as Ciaramelli and colleagues (2007) have demonstrated that patients with lesions to the VMPFC who have blunted affect but intact reasoning capacities also tend to make the same utilitarian judgments as normal participants on certain moral vignettes. Furthermore, the patients and the normal subjects displayed activation in the dorsolateral prefrontal cortex, like in Greene et al.'s above study. Since the patients have blunted affect and make the same utilitarian judgments as normal persons, this provides evidence that cognitive utilitarian reasoning is being used by normal agents in order to make normal moral judgments in certain cases. These studies show that what at least in part influences moral categorization in such cases is cognitive knowledge, not emotions. This is one important step for the establishment of the theory-theory since the theory view is a cognitivist one. Moreover, this cognitive knowledge involves a utilitarian calculation. In other words, it is knowledge of a master moral principle and is therefore, not prototype knowledge. Thus, this provides evidence for the theory-theory for moral concepts in that theory knowledge is being used at times in ordinary moral categorizations by normal participants just as theory knowledge of the function of a tuk was used by participants to judge that an object is not a tuk since the given object cannot perform the function of a tuk.

We now will turn to experiments that demonstrate that causal moral law knowledge at times influences moral categorization. This will provide additional independent evidence for the viability of the theory-theory for moral concepts. There is good evidence

---

moral judgment and deontological thinking, there is no causation; there is no general causal influence of deontological theory in moral categorization for normal subjects. However, even if Kahane is correct, then the trolley problems provide evidence for the use of deontological *theory* knowledge rather than utilitarian-like theory knowledge. Therefore, we still get our general desired conclusion, and the trolley experiments still provide evidence for the use of some kind of ethical theory knowledge in moral categorization. To note, Kahane likewise fails to draw the connection between moral psychology and the concepts literature in his writings.

that causal moral law knowledge is involved at times in moral categorizations (Cushman 2008; Young & Saxe 2008). As a moral categorization example, Paharia, et al. gave a group of subjects the following situation:

> A well-known real estate developer, X, owned a piece of property they wished to construct new housing units on. The property contained some health-threatening toxic substances that would require a substantial amount of clean-up, and was worth $50 million dollars as is. It would require $30 million to fully clean the land, but the value would only go up to $60 million. [The housing developer decided to only invest $12 million in a 40% clean up effort, and the value of the land went up to $54 million. They built housing units on the land, all of which have now been sold.] (Paharia 2009, 136)

Meanwhile, another group of participants received the same vignette except the text in the bracket was replaced with: "The housing developer sold the land to a lesser-known developer, Y, without cleanup. The lesser-known developer invested no money in any clean up effort and built housing units on the land, all of which have now been sold (136)." The experimenters discovered that participants judged developer X to be less unethical in the second situation when they sold the land to developer Y as compared to the first situation in which X directly rather than indirectly caused the property to not be fully cleaned up. Notice that they judged X to be less culpable in the situation where there was no clean up rather than in the case where there was a 40% clean up. Furthermore, in a third scenario, the experimenters found that these results held even when in later studies it was explicitly stated to subjects that agent Y was an instrument of agent X, contracted to do its bidding. These studies provide support for the theory-theory in that the best explanation for why these series of judgments are made in the three different scenarios is that participants take into account whether an agent is a direct or indirect cause for action (or inaction) when making moral categorizations. If only emotions such as anger or, for that matter, any kind of knowledge influences judgments without any conjoint influence whatsoever from the above causal knowledge regarding whether the agent is a direct or indirect cause of an action, then we should expect participants to claim that in all three situations, agent X is equally culpable or that perhaps X is more culpable in the second and third scenarios since there is no clean up whatsoever. However, the fact that subjects generally judge X to be most culpable in the first scenario, which is the only circumstance where X is directly responsible, strongly suggests that some kind of a general causal moral principle is in play and is being used in categorization in

order to properly explain and make sense of the discrepancy in judgments for the above three scenarios. This causal moral law knowledge places more culpability on those who directly causally influence an immoral outcome as contrasted with being an indirect cause of an immoral outcome. This causal principle knowledge is at work for many subjects when making certain judgments, regardless of whether emotions or some other kind of conceptual constituents are jointly also in play working together with the causal principle in moral cognition. That this knowledge is not an emotion is one important step in establishing the viability of the theory-theory for moral concepts since the theory view is a cognitive account of concepts. Furthermore, that this knowledge is of a causal moral law establishes that it is theory rather than prototype knowledge. Therefore, this study provides evidence for the use of theory knowledge in moral categorization.

Finally, Young, et al. discovered that causal moral law knowledge of an agent's causal intentions to act influences moral categorization (Young et al. 2010). In other words, an agent's causal intentions which influence the agent's behavior affect one's moral judgment of the agent. In this study, the experimenters used trancranial magnetic stimulation (TMS) in order to disrupt the neural activity in the right temporoparietal junction (RTPJ) before and during moral judgment. There has previously been shown to be a correlation between activation in the RTPJ and when a participant reads about an agent's causal intentions for action in certain moral contexts (Young et al. 2007). What they found was that in attempted harm cases, where the agent causally intends to do harm but fails to bring about the negative consequences, TMS to the RTPJ causes subjects to judge the agent's attempted harm as being less morally forbidden and more morally permissible as compared to participants who received TMS to a control site. Due to the nature of attempted harm cases, this suggests that TMS to the RTPJ affects the ability of subjects to fully account for the agent's causal intentions in making moral judgments and that assessing an agent's causal intentions does play a role in influencing typical normal moral judgments. This study demonstrates that in many normal moral judgments for normal subjects who are not receiving TMS to the RTPJ, such participants take into account causal knowledge about an agent's intentions and what causally motivates the agent to perform a given action. Normal participants have causal moral law knowledge that if an agent's causal motivations for action are bad, then the agent's action is still wrong even though the agent fails to bring about the bad consequences. This causal moral principle knowledge impacts their moral decision-making. Given that the causal knowledge of an agent's intentions for action affects the gravity of moral judgments, this provides evidence for the viability of the theory-theory for moral concepts in light of categorization. In this section, I have discussed several different sets of studies,

where each set can independently show that some people at least in part have theory knowledge stored in their moral concepts. My conclusion that some moral concepts have theory structure is an advancement beyond the current literature.

Conclusion

I purport to have provided two main contributions to the concepts literature in philosophy of cognitive science. I have first shown how the theory-theory will look like for moral concepts and have shown what kinds of knowledge are theory knowledge in the moral domain. Currently, the concepts literature does not state what counts as theory knowledge for moral concepts. Second, while I perfectly leave open the possibility that our moral concepts may also store different kinds of knowledge, such as prototypes, that are used individually or conjointly with other kinds of knowledge or information-carrying mental states in moral cognition depending upon the circumstances, I freshly have demonstrated that for moral categorization of what is morally right and wrong, there is in certain circumstances evidence for the theory-theory of moral concepts. Currently in the concepts field, it has not been stated nor explicitly shown that the theory view applies to the moral concepts domain for moral categorization. The relevant connection between the concepts and moral psychology fields has not been made in the literatures. While I perfectly leave open the possibility that theory knowledge may also apply to other different domains of concepts, my focus here is exclusively on moral concepts.

My conclusions are a non-trivial matter since as mentioned above, numerous experimental studies demonstrate that one cannot draw structural concept conclusions on abstract concepts based solely on data from concrete concepts. One cannot infer that moral concepts in part have theory structure based on the evidence for the theory-theory in the concrete concepts domain. Moreover, it is also non-trivial in that it has never been stated in the literature what theory knowledge even is for moral concepts; without which no one lucidly can claim that the theory view is viable for moral concepts. Furthermore, as we have discussed, several moral psychologists independently have had to devise and run numerous clever studies to sufficiently prove that the relevant knowledge is in fact used in moral cognition. The establishment of the theory view for moral concepts is hardly trivial.

A third contribution to the concepts field is that the above work on the theory view for moral concepts may open up new lines of future empirical work on explicitly examining what other kinds of theory knowledge may be stored in our moral concepts. For example, there are a variety of different ethical views in philosophy, such as

Aristotelian virtue ethics and Kantian deontological theory. Do some people store virtue ethical theory knowledge in their moral concepts or in part other theory knowledge? Also, while I have merely claimed here that *some* moral concepts have theory structure, there are many different types of moral concepts. More studies can be run on moral concepts not examined here, such as on HONESTY, INTEGRITY, and GREED, to see whether they also may have theory structure.

Although there presently is insufficient evidence to examine whether theory knowledge for moral concepts plays a role in other higher acts of cognition, such as in concept combination, induction, and analogical reasoning,[9] I have shown how experiments that were not explicitly designed to test for the theory-theory of moral concepts actually can be used to demonstrate that theory moral knowledge is used at times in moral categorization just as theory knowledge of hidden essences is used at times by many to classify an object as being a cow rather than being a horse. I leave further examination of the full extent of the use of theory knowledge in moral cognition for a later time. Recall that such theory knowledge can be used individually or conjointly with other kinds of knowledge in moral cognition depending on the context. There even may be situations where theory knowledge is not being used in particular cases of moral decision-making.

References

Barsalou, L., and K. Wiemer-Hastings. 2005. "Situating Abstract Concepts." In *Grounding Cognition*, edited by D. Pecher and R. Zwaan, 129–163. Cambridge: Cambridge University Press.

Carey, Susan. 1985. *Conceptual Change In Childhood*. Cambridge, MA: The MIT Press.

Carey, Susan. 2009. *The Origin of Concepts*. Oxford: Oxford University Press.

Ciaramelli, E., M. Muccioli, E. Ladavas, and G. di Pellgrino. 2007. "Selective Deficit in Personal Moral Judgment Following Damage to Ventromedial Prefrontal Cortex." *Social Cognitive and Affective Neuroscience* 2: 84–92.

Churchland, Paul. 1989. *A Neurocomputational Perspective*. Cambridge, MA: The MIT Press.

9. If moral concepts do use theory knowledge for categorization, we should expect that such knowledge should be able to partake in the other higher competences just as well as the theory knowledge in concrete concepts can.

Cushman, Fiery. 2008. "Crime and Punishment: Distinguishing the Roles of Causal and Intentional Analyses in Moral Judgment." *Cognition* 108: 353–380.

Fodor, J. 1998. *Concepts*. Oxford: Oxford University Press.

Frei, Jennifer, and Phillip Shaver. 2002. "Respect in Close Relationships: Prototype Definition, Self-Report Assessment, and Initial Correlates." *Personal Relationships* 9: 121–39.

Gopnik, Alison and Andrew N Meltzoff. 1997. *Words, Thoughts, and Theories*. Cambridge, MA: The MIT Press.

Greene, J, L.E. Nystrom, A.D. Engell, J.M. Darley, and J.D. Cohen. 2004. "The Neural Bases of Cognitive Conflict and Control in Moral Judgment." *Neuron* 44: 387–400.

Greene, Joshua, Sylvia Morelli, Kelly Lowenberg, Leigh Nystrom, and Jonathan Cohen. 2008. "Cognitive Load Selectively Interferes with Utilitarian Moral Judgment." *Cognition* 107*:* 1144–54.

Greene, Joshua. 2013. *Moral Tribes.* New York: The Penguin Press.

Haidt, Jonathan. 2012. *The Righteous Mind*. New York: Vintage Books.

Hampton, James. 1981. "An Investigation of the Nature of Abstract Concepts." *Memory & Cognition* 9 (2): 149–156.

Johnson, Mark. 1993. *Moral Imagination*. Chicago: The University of Chicago Press.

Kahane, Guy. 2012. "On the Wrong Track: Process and Content in Moral Psychology." *Mind & Language* 27: 519–545.

Keil, Frank C. 1989. *Concepts, Kinds, and Cognitive Development*. Cambridge, MA: The MIT Press.

Koenigs, M., L. Young, R. Adolphs, D. Tranel, F. Cushman, M. Hauser, A. Damasio. 2007. "Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgments." *Nature* 446: 908–911.

Lin, E. and G. Murphy. 1997. "The Effects of Background Knowledge on Object Categorization and Part Detection." *Journal of Experimental Psychology* 50A: 25–48.

Locke, John. (1689) 1996. *An Essay Concerning Human Understanding*. Edited by Kenneth P. Winkler. Indianapolis, IN: Hackett Publishing Company, Inc.

Machery, Edouard. 2009. *Doing Without Concepts*. Oxford: Oxford University Press.

Medin, D. and E. Shoben. 1988. "Context and Structure in Conceptual Combination." *Cognitive Psychology* 20: 158–90.

Mendez, M.F., E. Anderson, and J.S. Shapria. 2005. "An Investigation of Moral Judgment in Frontotemporal Dementia." *Cognitive and Behavioral Neurology* 18 (4): 193–7.

Mikhail, John. 2007. "Universal Moral Grammer: Theory, Evidence and the Future." *Trends in Cognitive Sciences* 11 (4): 143–152.

Murphy, Gregory. 2004. *The Big Book of Concepts*. Cambridge, MA: The MIT Press.

Murphy, G. and D. Medin. 1985. "The Role of Theories in Conceptual Coherence." *Psychological Review* 92: 289–316.

Paharia, N., K. Kassam, J. Greene, and M. Bazerman. 2009. "Dirty Work, Clean Hands: The Moral Psychology of Indirect Agency." *Organizational Behavior and Human Decision Processes* 109 (2): 134–141.

Park, John J. 2013. "Prototypes, Exemplars, and Theoretical & Applied Ethics." *Neuroethics* 6: 237–247.

Prinz, Jesse. 2002. *Furnishing the Mind*. Cambridge, MA: The MIT Press.

Quine, W.V.O. 1977. "Natural Kinds." In *Naming, Necessity, and Natural Kinds*, edited by S.P. Schwarz, 155-175. Ithaca, NY: Cornell University Press.

Rosch, Eleanor and Caroline Mervis. 1975. "Family Resemblances: Studies in the Internal Structure of Categories." *Cognitive Psychology* 7: 573–605.

Smith, Kyle, Seyda Smith, and John Christopher. 2007. "What Defines the Good Person?" *Journal of Cross-Cultural Psychology* 38: 333–360.

Snarey, John. 1985. "Cross-Cultural Universality of Social-Moral Development: A Critical Review of Kohlbergian Research." *Psychological Bulletin* 97: 202–32.

Stich, Stephen. 1993. "Moral Philosophy and Mental Representation." In *The Origin of Values*, edited by M. Hechter, L. Nadel, and R. Michod, 215–228. New York: Aldine de Gruyer.

Walker, Lawrence and Russell Pitts. 1998. "Naturalistic Conceptions of Moral Maturity." *Developmental Psychology* 34: 403–418.

Walker, Lawrence and Karl Hennig. 2004. "Differing Conceptions of Moral Exemplarity: Just, Brave, and Caring." *Journal of Personality and Social Psychology* 86: 629–647.

Wiemer-Hastings, K., and X. Xu. 2005. "Content Differences for Abstract and Concrete Concepts." *Cognitive Science* 29: 719–736.

Weiskopf, Daniel. 2009. "The Plurality of Concepts." *Synthese* 169: 145–173.

Wong, David B. 2006. *Natural Moralities*. Oxford: Oxford University Press.

Young, L., F. Cushman, M. Hauser, and R. Saxe. 2007. "The Neural Basis of the Interaction between Theory of Mind and Moral Judgment." *Proceedings of the National Academy of Sciences of the United States of America* 104: 8235–8240.

Young, L., and R. Saxe. 2008. "The Neural Basis of Belief Encoding and Integration in Moral Judgment." *NeuroImage* 40 (4): 1912-1920.

Young, L., J. Camprodon, M. Hauser, A. Pascual-Leone, and R. Saxe. 2010. "Disruption of the Right Temporoparietal Junction with Transcranial Magnetic Stimulation Reduces the Role of Beliefs in Moral Judgments." *PNAS* 107 (15): 6753–6758.

cognethic.org