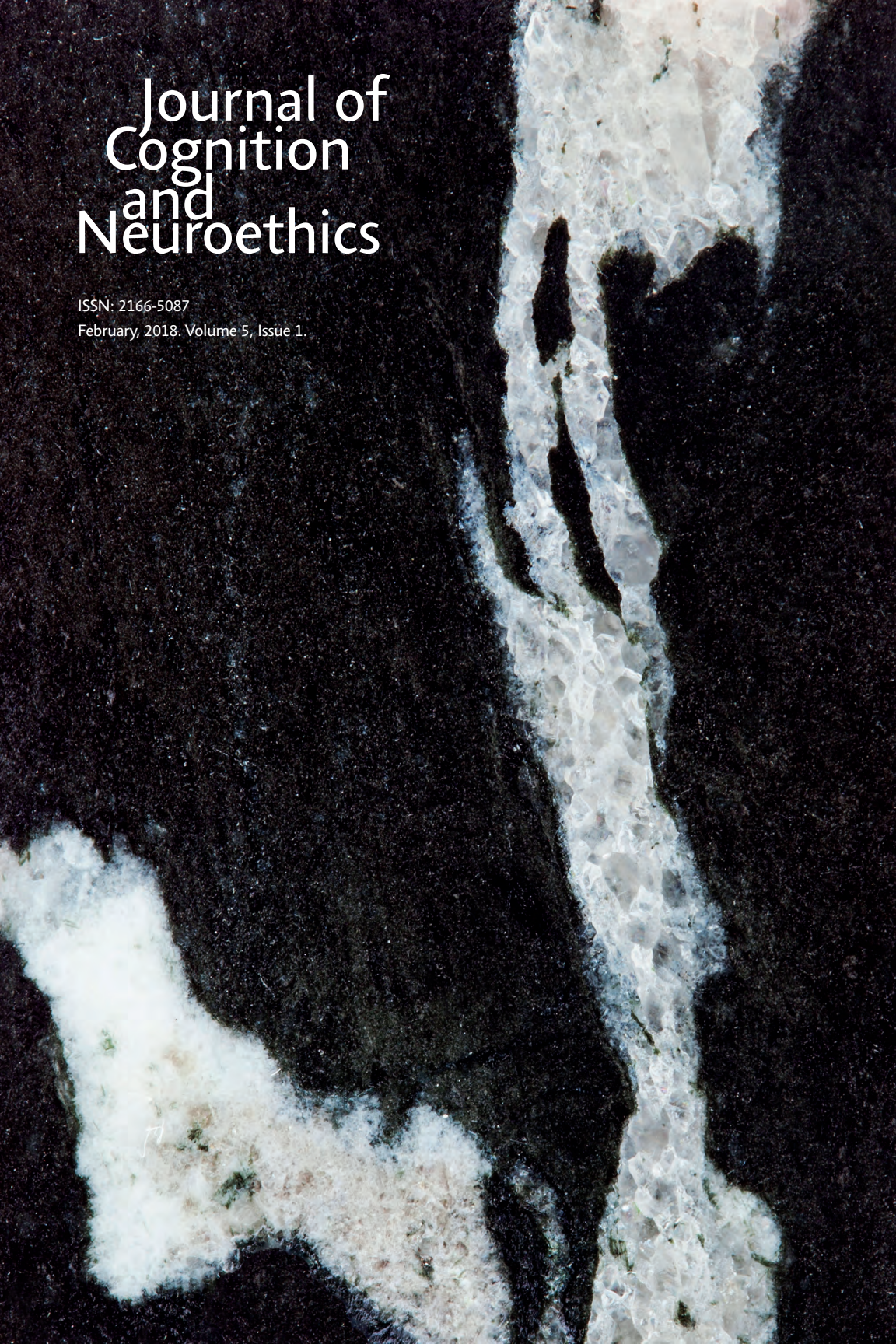
Journal of Cognition and Neuroethics

ISSN: 2166-5087
February, 2018. Volume 5, Issue 1.



Journal of Cognition and Neuroethics

Managing Editor
Jami L. Anderson

Production Editor
Zea Miller

Publication Details
Volume 5, Issue 1 was digitally published in February of 2018 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

Phantom Sensations: What's a Brain to Do? A Critical Review of the Re-mapping Hypothesis

Daniel J DeFranco
Tulane University

Biography

I am doctoral candidate in Philosophy at Tulane University, and will be defending my dissertation Spring 2018. I work in the areas of Early Modern Philosophy, Philosophy of Mind, and Neuroscience, and have a particular interest in exploring the connections between the history of philosophy and contemporary neuroscience.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). February, 2018. Volume 5, Issue 1.

Citation

DeFranco, Daniel J. 2018. "Phantom Sensations: What's a Brain to Do? A Critical Review of the Re-mapping Hypothesis." *Journal of Cognition and Neuroethics* 5 (1): 1–25.

Phantom Sensations: What's a Brain to Do? A Critical Review of the Re-mapping Hypothesis

Daniel J DeFranco

Abstract

Here, I will review the most widely held account of phantom sensations and the effects of deafferentation in the somatosensory cortex, namely, the "re-mapping hypothesis." According to the re-mapping hypothesis, deafferentation is followed by significant neural reorganization that eliminates the neural structures that give rise to phantom sensations, restoring the alignment between the brain's representation of the body and the actual condition of the body. Implicit in the re-mapping hypothesis is the view that the brain's primary function is the accurate representation of the body. In response to the remapping hypothesis, I propose an alternative theory, which I have dubbed the "preservation hypothesis." The preservation hypothesis argues that the primary function of the brain is to preserve the entirety of the brain's structures and functional capacities. Thus, upon deafferentation, the brain does not eliminate phantom sensations and restore an accurate representation of the body but takes steps to preserve the neural structures underlying phantom sensations, with the effect of maintaining phantom sensations long term. While the effects of deafferentation are certainly an empirical matter, assessing our views on the subject discloses our deeply held assumptions regarding the primary function of the brain: does the brain operate such that it will do all that it can to represent the body accurately? Does the brain have certain limitations in its accurate representation of the body? Or, does the brain care nothing for reality and the accurate representation of the body, operating with the sole purpose of preserving its structure and functional capacities in their entirety? I hope to make some progress in answering these questions.

Keywords

Phantom Sensations, Brain Function, Remapping, Preservation

Whenever I tell someone that I am researching phantom limbs,[1] I often get the same question in response, "why doesn't the brain know that the limb is no longer there?" While the question may be posed by a novice, implicit in the question is a sophisticated understanding of the function of the brain. To ask, "why doesn't the brain know that the limb is no longer there," implies that the brain's production of an experience of the body that does not align with the actual condition of the body constitutes a failing on the part of the brain. Interestingly, it is not just laymen who hold this view; neuroscientists

1. The phantom limb, the lingering feeling that one's amputated limb is still present, is a condition experienced by 98% of amputees, 60-80% of whom experience some degree of pain associated with the phantom (Ramachandran and Hirstein, 1998; Sherman *et al*. 1984).

also inquire, "what happens in the brain upon deafferentation?"[2] That is, does anything happen in the brain after amputation that might correct the subsequent discordance between the experience and actual condition of the body?

In what follows, I will critically evaluate both current explanations for phantom sensations and the conceptions of the brain to which these explanations are implicitly committed. I will begin with a review of the most widely-held account of phantom sensations, the "re-mapping hypothesis." Implicit in the re-mapping hypothesis is the belief that the mechanics of the brain work to represent the body accurately, and, upon amputation, significant neural reorganization is initiated to eliminate the causes of phantom sensations and restore the alignment between experience and actual condition of the body. I will assess this hypothesis, raising issues with it, and ultimately proposing an alternative account, which I have dubbed the "preservation hypothesis." The preservation hypothesis posits that the brain does not attempt to eliminate phantom sensations and accurately represent the body but takes steps to preserve the neural structures underlying phantom sensations, with the effect of the long-term maintenance of phantom sensations.

While the causes of phantom sensations and the effects of deafferentation are certainly empirical matters, assessing our views on the subject discloses our deeply held assumptions regarding the ultimate function of the brain: does the brain operate such that it will do all that it can to represent the body accurately? Does the brain have certain functional limitations in the accurate representation of the body? Or, does the brain care nothing for reality and the accurate representation of the body, operating with the sole purpose of preserving its structures and functional capacities in their entirety? In this essay, I hope to approach answers to these questions.

I. The Initial Appearance of Phantom Sensations

With the aid of modern science, the initial appearance of phantom sensations is well understood. In the parietal lobe, exists an area of the brain called the somatosensory cortex (the primary sensory cortex), which is responsible for mapping sensations of the most peripheral parts of our body (i.e., the sensations of the skin, joints, and muscles). Whenever we touch something, or something touches a part of our body, afferent signals via the peripheral nervous system are sent to the part of the somatosensory cortex that corresponds with that limb, those afferent signals are mapped and we experience a

2. Deafferentation is a disruption of afferent neural connections between the peripheral body and brain. In this piece, I will be considering deafferentation that results from the amputation of a limb.

sensation in the limb. If the nerves connecting a body part to the brain were disconnected or destroyed, or if a body part were amputated (i.e., deafferented), the part of the somatosensory cortex representing that part of the brain would continue to exist, at least for a time, despite not receiving any direct afferent signals from the body. Additionally, the deafferented cortical area continues to receive efferent signals from the motor cortex, signals concerning executive movements of the body. The continued existence of a deafferented cortex's neural structure, in conjunction with its continued stimulation by efferent signals from the motor cortex gives rise to phantom sensations and explains the initial appearance of the phantom limb. The arrival of efferent signals from the motor cortex also explains why so many amputees claim that they can move their phantom.[3]

Many mysteries continue to surround phantom sensations, and one of these mysteries proves almost as unusual as the case of the phantom limb itself. In 1991, Pons *et al*. made an incredible discovery while recording neuronal activity in the primary somatosensory cortex of four adult macaque monkeys. The macaques "had received deafferentations of an upper limb, three unilateral and one bilateral," by way of dorsal rhizotomy, "more than 12 years before the recording session" (Pons *et al,* 1857).[4] Pons noticed that when the faces of his monkey subjects were touched, the effect in the brain was such that the corresponding face area of the somatosensory cortex was activated (which was expected), but so was the area that represented the deafferented limb (Pons *et al*. 1991). Touching the ipsilateral face of a monkey with unilateral arm deafferentation excites the area of the somatosensory cortex that represents the face as well as the area that represents the arm. According to Pons *et al*. "[V]irtually identical findings were obtained in the three other animals," and just "a slight deflection of facial hairs was sufficient to obtain a vigorous neuronal response" in the deafferented zone (Ibid, 1859).

Prior to Pons' discovery, all attempts at researching phantom sensations faced the seemingly insurmountable obstacle of determining how to research a phenomenon

3. Studies by Lotze *et al*. 1999, Raffin, E. *et al*. 2012, and Makin *et al*. 2013 highlight amputees' motor control over their phantom limbs, many of whom can execute fine motor skills even 53 years after amputation. Motor skills executed by amputees include, but are not limited to, (1) elbow flexion/extension (for above-elbow amputees only); (2) wrist flexion/extension; (3) hand closing/ opening; (4) thumb to index opposition; (5) finger abduction/ adduction (Raffin, E. *et al*. 2012, 748).

4. Pons' study on the "Silver Spring monkeys" was the focus of one of the most public animal abuse cases in the history of the United States, and was the very first animal research case to reach the Supreme Court. Since Pons, phantom limb research has overwhelmingly utilized non-invasive and humane procedures on human patients. This change is in part due to the pioneering work and nuanced approaches to the study of phantom sensations by V.S. Ramachandran.

that had no physical correlate on the peripheral body. Pons *et al.*'s discovery (that the stimulation of intact body parts can excite a deafferented area in the somatosensory cortex) opened the possibility of studying phantom sensations through intact limbs albeit not the limb that corresponds to the phantom sensation. The discovery also raised a multitude of questions that researchers, thanks to Pons *et al.*, now had a means of empirically investigating. For instance, what was being observed in the co-activation of the face and hand cortices when the ipsilateral face of a monkey with unilateral hand deafferentation was stimulated? Additionally, what were the perceptual correlates of this experiment, that is, what were the monkeys feeling when their faces were touched?

Unfortunately, Pons could not ask the monkeys what they felt when their faces were touched, but, the following year, V.S. Ramachandran conducted an experiment with human unilateral hand amputees in which he could ask his patients what they felt when their faces were stimulated. Results showed that the perceptual correlates of this cortical co-activation are the simultaneous experience of sensations on the face and phantom hand (Ramachandran *et al.* 1992). These sensations are: (1) modality specific (i.e., if cold water is dripped on the face, then the patient will feel cold water being dripped on the phantom hand); and, (2) topographically organized (i.e., a point-by-point map of the surface of the hand appears on the face) (Ramachandran *et al.* 1992). Ramachandran's MEG scans of the somatosensory cortex confirm Pons *et al.*'s findings; when the ipsilateral face area of the somatosensory cortex is stimulated, both the face area and large sections of the deafferented hand area are simultaneously stimulated in the brain.

But what could explain the cortical co-activation that both Pons and Ramachandran were observing? As a first step in explaining this phenomenon, Ramachandran notes that the face area neighbors the hand area in the somatosensory cortex. And while the proximity of these two areas does not fully explain observed neural activity nor its perceptual correlates, it offers some insight into why touching the ipsilateral face of a unilateral hand amputee could simultaneously produce sensations in the face and the phantom hand. Unilateral hand amputees can also experience sensations in their phantom limbs when the arm area most proximal to the amputation line is stimulated, which is an area that also neighbors the hand area in the somatosensory cortex (Ramachandran *et al.* 1992). Additionally, stimulation of the contralateral intact hand also elicits sensations in the deafferented hand area; here we are observing "cross-callosal" connections as opposed to intra-hemispheric connections between neighboring areas of the somatosensory cortex (Giummarra *et al*. 2007).

II. The Remapping Hypothesis

As an explanation of the co-activation of intact and deafferented sensory areas, Ramachandran formulates the "Remapping Hypothesis," which remains to this day the most widely accepted account of cortical co-activation. The remapping hypothesis proposes the following:

1. Upon deafferentation, a cortical area becomes vacant or empty, for no direct sensory input is arriving to this part of the brain from the peripheral body.

2. Neighboring cortical areas sense the vacancy in the deafferented cortex, and actively invade with the effect of reorganizing the deafferented cortical area such that its structure and functional capacities become continuous with the intact cortical area.

As for the mechanism that facilitates this cortical reorganization, Ramachandran offers the following two possibilities:

1. The invasion of a deafferented cortical area occurs when a neighboring cortical area "sprouts thousands of neural tendrils that creep over into" the deafferented area (Ramachandran 2011, 28).

2. Preexisting neural connections exist between neighboring cortical areas, which are "masked" or "inhibited" under normal conditions (i.e., when a cortex is receiving afferent signals from the peripheral body) (Ramachandran *et al.* 1992, 1160; Ramachandran 2011, 28). Upon deafferentation, a cortical area is no longer able to inhibit signals from these pre-existing neural connections, resulting in an invasion of afferent signals from neighboring cortical areas.[5]

5. Ramachandran explains, "Thus even in healthy normal adult brains there might be sensory inputs from the face to the brain's face map *and* to the hand map area as well. If so, we must assume that this occult or hidden input is ordinarily inhibited by the sensory fibers arriving from the real hand. But when the hand is removed, this silent input originating from the skin on the face is unmasked and allowed to express itself so that touching the face now activates the hand area and leads to sensations in the phantom hand" (Ramachandran 1998, 34).

Considering that phantom sensations can develop quite quickly post-amputation,[6] Ramachandran suggests that the *initial* work of cortical reorganization is most likely "a result of the unmasking of 'silent' synapses, rather than of anatomical changes, such as 'sprouting'" (Ramachandran *et al*. 1992, 1160). However, the two mechanisms of cortical reorganization are not mutually exclusive. Ramachandran explains that there is presently "no way…of easily distinguishing between these two theories, although my hunch is that both mechanisms are at work." (Ramachandran 1998, 35). Thus, cortical reorganization could very well be initiated by the unmasking of pre-existing synaptic connections, while the long-term process of reorganization involves the sprouting of new neural connections.

Ramachandran also surmises that this process is evolutionarily "beneficial to the organism" (Ramachandran 2000, 319). For one, the process restores the alignment between the experience of the body and the actual condition of the body by eliminating the neural structures that give rise to phantom sensations. Ramachandran views the restoration of the alignment between experience and the condition of the body as the mechanism's primary function. In addition to restoring the discordance between experience and actual condition of the body, cortical reorganization prevents cortical degeneration. Without sufficient afferent stimulation, a cortical area would degenerate and ultimately become a rotted-out bit of brain. The neural mechanism at work post-deafferentation prevents this degeneration; an intact cortical area appropriates a deafferented area and includes it in its own representational activities, guaranteeing continued afferent stimulation and thus preventing neural degeneration in the somatosensory cortex (Ramachandran 2000, 319).

Following the appropriation of a deafferented cortex by an intact cortex, Ramachandran suggests that we can reasonably anticipate an increase in that intact cortex's representational and functional powers (Ramachandran 2000, 319). An intact cortex would come to possess a greater area of representation in the somatosensory cortex, and there ought to be measurable and perceptual correlates to this change. Specifically, Ramachandran anticipates "tactile hyperacuity" both in the area most proximal to the amputation line and in any body parts whose area of correspondence in the brain neighbors the deafferented area.

6. Patients can experience vivid phantom sensations immediately following the amputation of a limb, such that "[S]ome patients wake up from anesthesia and are incredulous when told that their arm had to be sacrificed, because they still vividly *feel* its presence" (Ramachandran 1998, 22).

Beyond its ability to explain cortical co-activation, the remapping hypothesis proves fascinating, insofar as it implicitly contains a vision for the primary function, the telos, of the brain. According to Ramachandran, deafferentation initiates a mechanism in the somatosensory cortex by which intact cortical areas invade and reorganize a deafferented cortex with the effect that the deafferented cortex is completely appropriated into intact areas. The perceptual effect of this mechanism is that phantom sensations, over time, decrease in their vividness and eventual vanish completely, restoring the alignment between one's experience of the body and the actual condition of the body. And if this mechanism is indicative of an overall function of the brain, then it seems clear that the brain's utmost concern is producing an accurate representation of the body. Should something challenge the brain's ability to accurately represent the body, such as deafferentation, the brain initiates a mechanism by which it corrects the discrepancy. In the case of phantom sensations, the brain must literally destroy a part of its structure and functional capacities in order to represent the body as it actually is.

The principle question asked by neuroscientists with respect to phantom sensations is, "what are the effects of deafferentation in the brain" (i.e., what happens in the brain after a limb is amputated?). This question proves to be no more than a sophisticated form of the same question posed by non-experts on learning of phantom sensations, which is, "why doesn't the brain know that the limb is no longer there?" The intuitions of experts and laypeople are aligned when it comes to phantom sensations; any condition in which the experience of the body is in obvious discord with the actual condition of the body is an intolerable one. Either the brain must do something to correct this condition or the brain has hit a functional limit in its ability to accurately represent the body. With respect to phantom sensations, the remapping hypothesis takes the position that the brain can come to know that the limb is no longer present, and initiates a mechanism of neural reorganization to eliminate the neural structures that give rise to phantom sensations, restoring the alignment between the experience and condition of the body. Guiding this hypothesis, and the question posed by experts and laypeople alike, is the intuition that the primary function of the brain is to accurately represent the condition of the body. If the brain fails to accurately represent the body, as in the case of phantom sensations, then the brain must do something to correct the situation or the error signals an instance in which the brain has been stumped (i.e., has hit a functional limit in its ability to accurately represent the body). Regardless of what view is taken, both views are equally guided by the assumption that brain's primary function is the accurate representation of the body. Any discordance that arises between the experience and actual condition of the body must be corrected or constitutes a failure of the brain.

III. A Cartesian Interlude

Nearly 400 years ago, another novel scientist took up the challenge of explaining phantom sensations. This scientist determined that sensations must be occurring in the brain, not the body, and that the continued existence of structures in the brain that directly correspond to the peripheral body support the emergence of phantom sensations; "each tiny tube on the inside surface of the brain corresponds to a bodily part" and "each point on the surface of gland H [the pineal gland] corresponds to a direction in which these parts can be turned" (Descartes *The World*, 154). And just as Ramachandran's explanation of phantom sensations contains within it a vision of the primary function of the brain, this scientist also understood phantom sensations as offering insight into the function of the brain, a window into the teleology of mind. His name: Descartes.

Phantom sensations are often viewed as the exclusive domain of contemporary science, and while some researchers have explored the field's early modern origins,[7] most insights remain in the annals of history, failing to transcend into the contemporary scientific discussion. Here, I would like to indulge in an experiment. I will review Descartes' treatment of phantom sensations, not as a documentation of the past, but as inspiration for assessing the state of phantom limb research as it stands today.

In Meditation 6 of *Meditations on First Philosophy*, Descartes explains the phenomenon of phantom sensations as follows,

> …many experiences gradually weakened any faith that I had in the senses…And not just the external senses, but the internal senses as well. For what could be more intimate than pain? But I had heard it said by people whose leg or arm had been amputated that it seemed to them that they still occasionally sensed pain in the very limb they had lost. Thus, even in my own case it did not seem to be entirely certain that some bodily member was causing me pain, even though I did sense pain in it (Descartes *Meditations*, 95).[8]

7. Finger and Hustwit offer an extensive account of the historical development of phantom limb research in their article "Five Early Accounts of Phantom Limb in Context: Pare, Descartes, Lemos, Bell, and Mitchell" (Finger and Hustwit 2003).

8. In addition to Meditation 6 of the *Meditations on First Philosophy*, Descartes discusses phantom sensations on at least two other occasions; in a private letter to Fromondus (1637) and in Part IV Article 196 of *Principles of Philosophy* (1644).

Phantom sensations are treated by Descartes on two different levels of understanding. The first level, the common-sense level of understanding, treats phantom sensations with respect to how they initially strike us as a failing of the body/brain. Phantom sensations lead one to doubt the reliability of the senses, because, in losing a limb, the senses fail to represent the condition of the body accurately. Phantom sensations are to feel what is not there and to know that which is false. This common-sense level of understanding understands phantom sensations themselves to be the error committed by the body/brain, an error in the accurate representation of the body.

The second level, the higher level of understanding, emerges from further reflection on the exact nature of the error to which phantom sensations alert us. As Descartes notes, phantom sensations raise the possibility that "even in my own case it did not seem to be entirely certain that some bodily member was causing me pain, even though I did sense pain there" (Ibid). Phantom sensations led Descartes to reconsider the location in the body/brain at which sensations are produced. Sensations are experienced as occurring in the body, but if sensations continue to be perceived in a limb even after that limb has been destroyed, then clearly the limb itself cannot be the locus of sensation. And if all sensations are experienced as occurring in the body, despite not actually occurring in the body, then the nature of sensations generally would be deceptive.

Descartes ultimately determines that sensations occur not in the body, but in the brain. Descartes explains,

> when nerves in the foot are agitated in a violent and unusual manner, this motion of theirs extends through the marrow of the spine to the inner reaches of the brain, where it gives the mind the sign to sense something, namely, the pain as if it is occurring in the foot (Ibid, 102).

Movements in the peripheral body travel up into the interior of the brain via nerves and the "marrow of the spine," but it is the movement in the brain that "gives the mind the sign to sense something." And there seems nothing inherently strange about sensations occurring in an organism's brain as opposed to its body.

It is unnerving, however, that the sensations the brain produces are experienced as occurring in the body as opposed to being experienced in their true location of occurrence in the brain. As Descartes notes,

> …the nature of man could have been so constituted by God that this same motion in the brain might have indicated something else to the mind: for example, either the motion itself as it occurs in the brain, or

> in the foot, or in some place in between, or something entirely different (Ibid).

The mechanics of the brain need not have been organized such that the production of sensation be experienced by an agent as occurring in that agent's body. Sensation could have been experienced, Descartes posits, as: (1) directly occurring in the brain; (2) as the actual motions in the body (i.e., violent agitation, as opposed to the feeling of pain); (3) as traveling through the nerves or up the spinal cord; or, (4) something entirely different. All these alternative possibilities posit ways in which sensations could have provided objective insights into the nature of sensation and its actual process of production. However, sensation, as it is, communicates something false about the nature of sensation; sensation is produced in the brain but is deceptively experienced as occurring in the peripheral body.

At a higher level of understanding, we see that phantom sensations do not independently constitute an error of sensation; rather, they alert us to an error that concerns the status of all our sensations. Sensations generally, insofar as they occur in the brain but are experienced in the body, fail to accurately communicate to us the actual happenings of the body. And this illusion, that sensations occur in the body when they in fact occur in the brain, is a direct product of the mechanics of the brain. We are no longer dealing with an error, but a failing, an infirmity of the brain itself. It was assumed that the brain's job is to accurately represent the body and the external world, but we find that the mechanics of the brain produce sensations that are deceptive with respect to sensation's actual nature and its true location of production. Sensation, as a potential mechanism for discerning the objective properties of the external world, is corrupt at its core.

Speaking for myself, I am rather grateful to my brain for failing to represent sensation as occurring at its true location of inception, and I cannot begin to imagine what it would be like to experience all the sensations we associate with our bodies as being experienced in the brain. For those of you who also intuitively grasp the benefit of experiencing sensations as occurring in our extended body, one may begin to see how Descartes has rather cheekily demonstrated that neither sensation nor, consequently, the mechanics of the brain are particularly interested in accurately representing the body and the happenings of the body. While the nature of man could have certainly been constituted differently, "nothing else would have served so well the maintenance of the body" than its present arrangement (Ibid). Thereby, for Descartes, the brain's disinterest in representing the body accurately is to our benefit; it is far better for our preservation

that we experience ourselves as a unified embodied being, and not as a being within a being (i.e., a smaller organism operating out of the head of a larger organism).

In reviewing the nature of sensation, Descartes uncovers two critical insights. Firstly, the nature of sensation as inherently deceptive provides grounds by which Descartes can reject his previously held assumption that the function of the body/brain is the accurate representation of the body. Secondly, Descartes recognizes that, although sensations are deceptive, their nature is such that "nothing else would have served so well the maintenance of the body" (Ibid, 102). Thereby, the true function of the body/brain, that which all the activities of the body/brain are directed towards, must be "the welfare of the body" (Ibid, 103).

There is one last hurdle Descartes must overcome in positing this new function for the body/brain; phantom sensations themselves. It is quite possible that phantom sensations constitute a moment in which the mechanics of the brain have failed to secure what is necessary to preserve the body/brain. It is to our benefit that the mechanics of the brain produce the experience of sensation as occurring in our body, but it is a burden that these same mechanics maintain the experience of a limb even after that limb has been destroyed. According to Descartes, "the nature of man…cannot help being sometimes mistaken," but "I know that all the senses set forth what is true more frequently than what is false regarding what concerns the welfare of the body" (Ibid, 102–3). While Descartes does not muse much over how phantom sensations could themselves be beneficial to the organism, he does believe that, regardless of whether phantom sensations are a benefit or a burden, the burden is absolutely worth the benefit.[9] That is: (1) to experience sensations as occurring in the body as opposed to in the brain; and, (2) to have the structural and functional capacities of the body enshrined in the brain such that sensations of the body arise in the brain, is far more beneficial for the welfare of the human than the burden posed by phantom sensations.

The burden, phantom sensations, does offer a conciliation prize, for a reflection upon phantom sensations aid us in further fleshing out the primary function of the brain and mind; the maintenance or preservation of the organism (Ibid 102-3). According to Descartes, "I can think of no better arrangement" than the current mechanics of the

9. Giummarra *et al*. argue that the preservation, and not the reorganization, of a deafferented cortex is necessary for the well-being of the individual. For instance, the successful operation of a prosthesis depends upon that prosthesis having a "neural template" in the somatosensory cortex. Without the preservation of a deafferented cortex, it would be impossible for a patient to operate or even recognize a prosthesis as an extension of his or her body (Giummarra *et al*. 2007, 223).

brain and the kind of sensations that it produces, for the sensation that it does produce, "of all the ones it is able to produce, is most especially and most often conducive to the maintenance of a healthy man" (Ibid, 102). Descartes is unequivocal in "Meditation 6" that the aim of brain mechanics is to produce a condition that best enables the human to maintain its health and preserve itself. And phantom sensations demonstrate a facet of the brain's function. Specifically, phantom sensations show us that the effort to preserve extends to the brain itself, where the brain preserves its structure and functional capacities even if a part of the body is destroyed. And the fact that the brain maintains its structure and functional capacities in spite of the actual condition of the body, demonstrates to Descartes that the function of preservation is not aligned with the function of accurate representation of the body and external world. The true function, or telos, of the mechanics of the brain is the maintenance and preservation of the entirety of the organism.[10]

At a common-sense level of understanding, phantom sensations themselves represent a failing of the brain to produce an experience of the body that accurately represents the body's present condition. At a higher level of understanding, we see that this failing is no failing at all, but an indication that our initial conception of the function of the brain is false. The function of the brain is the preservation of the organism's structure and functional capacities, and this function involves an indifference to the accurate representation of the condition of the body and the objective nature of the external world. Sensations are experienced as occurring in the body, despite originating in the mind, because this kind of sensation is most beneficial to the maintenance of the human organism. And while phantom sensations may be a burden to an amputee, the brain's preservation of its ability to represent the structure and functional capacities of a limb, even after that limb has been destroyed, offers a privileged glimpse into what follows from the brain's fulfilling its function of maintenance and preservation.

Returning to our contemporary study of phantom sensations, we see that Ramachandran certainly appreciates the common-sense understanding of phantom sensations, but perhaps has not grasped the higher sense understanding. Ramachandran recognizes that the phenomenon of phantom sensations constitutes a disconnect between our experience of the body and the actual condition of the body, but simultaneously maintains that the brain has a mechanism by which it can correct this disconnect and

10. In his book *Cartesian Metaphysics and the Whole Nature of* Man, Richard Hassing reiterates the point that "Descartes makes clear that the soul-body composite has a natural teleology: its natural end is the health and conservation on the body" (Hassing 2015, 57).

realign our experience of the body with the body's physical condition. By way of this mechanism, the brain can fulfill its function of accurately representing the body.

Ramachandran's diagnosis and suggested remedy of the problem, however, fails to recognize that the original sin is not phantom sensations themselves, but that sensations generally are deceptive; sensations occur in the brain but are experienced as occurring in the body. While this phenomenal aspect of sensations might be incredibly beneficial to the organism, it contradicts our account of the brain as working toward providing an accurate representation of our body and the external world. The brain does not only just fail to accurately represent the condition of the body post-amputation but it also fails to do so with respect to all sensations.

The illusion that Ramachandran simultaneously falls prey to and maintains is that there ever was a point in time at which the mechanism of the brain produced sensations that accurately communicated the condition of the body. Sensations, at their core, are deceptive, insofar as they do not accurately communicate the process of their production and the location of their inception, and thus have never satisfied the function of accurately representing the condition of the body. And, if that is the case, then it would be odd that, following deafferentation and the emergence of phantom sensations, the brain would suddenly take up an interest with the accurate representation of the body and initiate a rather dramatic process of neural reorganization such that the experience of the body is realigned with the actual condition of the body.

IV. The Preservation Hypothesis

The intuition that overwhelmingly guides the remapping hypothesis is that the primary function of the brain is the accurate representation of the body. Descartes, however, makes a persuasive case that sensations have never constituted an accurate representation of the body, and, therefore, it would be misguided to interpret neural changes following deafferentation as indicative of a restoration of accurate representation. How can something that never was be restored? This suggests that the remapping hypothesis could be correct in its observations of phantom sensations, but misguided in its assessment of the significance of these observations. Descartes' point warrants a re-examination of the evidence that ostensibly substantiates the remapping hypothesis.

The remapping hypothesis, in large part, depends upon evidence of cortical co-activation (i.e., the stimulation of a deafferented cortex by way of afferent signals overflowing from a neighboring intact cortex). Traditionally, cortical co-activation has been understood to indicate dramatic cortical reorganization, but it is critical to point out

that we do not literally see cortical re-organization in neural scans. Neural scans indicate where in the brain cortical activity is occurring, but do not show structural changes. In amputees, scans show activity in a deafferented cortex when there is activity in a neighboring intact cortex. Ramachandran takes this co-activation to indicate that an intact cortex has initiated an invasion of a deafferented cortex, such that the deafferented cortex is being reorganized and will become continuous with the intact cortex. But certainly, this cannot be the only possible explanation of cortical co-activation. It's possible that the co-activation of a deafferented and intact cortex has no significant effect at all. It is also possible that the co-activation of a deafferented and intact cortex serves to preserve the structure and functional capacities of the deafferented cortex.

Without afferent stimuli a cortex will degenerate, but the observed phenomenon of cortical co-activation indicates a way in which a cortex could continue to receive necessary afferent stimulation and avoid degeneration post-deafferentation. However, it is unclear if the origin of afferent signals affect the effects these signals can have on a deafferented area. That is, must the afferent stimuli arriving at a cortex originate from that cortex's corresponding limb to contribute to its preservation? If the answer is yes, then it seems doubtful that the effect of cortical co-activation is the preservation of a deafferented cortex. But if the answer is no, that afferent signals can stimulate a cortex regardless of their location of origin, then the afferent stimuli of diverse origins could very well serve to preserve the structure and functional capacities of a deafferented cortex.

Fortunately, progress has been made in answering the above question. In an extensive review of the phantom limb literature, Giummarra *et al*. argue that afferent stimuli continue to arrive at a deafferented limb from a variety of sources and that this continued afferent stimuli has the effect of preserving the neural structure of the deafferented area, giving rise to a "normal (non-painful)" phantom limb (Giummarra *et al*. 2007, 228). A deafferented cortex's sources of afferent stimulation include: (1) "the residual limb and stump;" (2) sensations arriving from the intact contralateral limb via "cross-callosal pathways; (3) "activation of mirror neurons" from watching others move their intact limbs; and, (4) visual feedback from the use of a prosthetic (Ibid, 224; 226; 225; 223). Giummarra *et al.* argue that the arrival of afferent stimuli from these diverse sources counteract cortical reorganization initiated by neighboring intact cortices.[11] Thus,

11. This position is also echoed by Lotze *et al*, who argue that "frequent and extensive use of a myoelectric prosthesis is correlated negatively with cortical reorganization and phantom limb pain and positively with the reduction in phantom limb pain over time. This suggests that the ongoing stimulation, muscular training of the stump and visual feedback from the prosthesis might have a beneficial effect on both

afferent stimuli arriving at a deafferented cortex can aid in preserving the structure and functional capacities of that cortex despite not originating from the peripheral limb to which the cortex corresponds.

Now, if afferent signals coming from other sources have the effect of stabilizing the structure and functional capacities of a deafferented cortex, why must afferent signals originating in neighboring intact cortices destabilize the structure of a deafferented cortex, as suggested by the remapping hypothesis? It is certainly possible that Ramachandran is correct, that afferent signals from an intact cortex have the effect of reorganizing a deafferented cortex, but such a position requires further explanation; we must know what accounts for the different effects had by afferent stimuli in a deafferented cortex.

There is, of course, the possibility that afferent stimuli arriving from intact cortices serve to preserve a neighboring deafferented cortex. And, in fact, the findings of Giummarra *et al*. substantiate the view that cortical co-activation constitutes a mechanism by which the structure and functional capacities of a deafferented cortex are preserved long-term. I call this alternative account of cortical co-activation the "preservation hypothesis," and I summarize it as follows:

1. Upon deafferentation, the brain's normal mode of operation is maintained, and the effect is the preservation of the structure and functional capacities of the deafferented area.

2. Preservation is accomplished by an overflow of afferent signals from intact cortical areas arriving at and stimulating the deafferented area.

3. The arrival of these afferent signals is made possible by pre-existing neural connections that exist between the deafferented area and many other cortical areas.

4. The flow of afferent signals does not represent a significant change in cortical organization or operation, but reveals the normal functioning and communication between brain regions.[12]

---

cortical reorganization and phantom limb pain" (Lotze *et al*. 1999, 502).

12. Afferent signals normally overflow into other cortical areas, but usually are inhibited by afferent signals coming from intact peripheral limbs (Ramachandran 1998, 34).

The preservation hypothesis proposes that no significant cortical reorganization occurs in a cortical area following deafferentation. Rather, the brain's structures and modes of operation facilitate the stimulation of a deafferented area with afferent stimuli from intact cortical areas, the effect of which is that the deafferented area receives stimuli aiding it in preserving its overall structure and functional capacities. In this view, a cortical area would have many pre-existing neural connections between itself and a variety of other cortices, receiving continuous afferent stimulation. Under normal operating conditions, these neural signals are inhibited by afferent signals arriving directly from the peripheral body. When the arrival of afferent signals directly from the peripheral body desists following deafferentation, afferent signals coming from other cortices – signals that have been arriving the entire time but until this point have failed to excite the intact cortical area – now successfully excite the deafferented area. The arrival of these afferent signals serve to stimulate the deafferented cortex with the effect of preserving the deafferented cortex's structure and functional capacities. This is not to say that afferent signals coming from other cortical areas are sufficient for maintaining the structure and functional capacities of the deafferented cortex, but that this is the end that the afferent signals serve.

Note that the preservation hypothesis utilizes a great deal of Ramachandran's initial insights regarding phantom sensations and the effects of deafferentation. The point of disagreement concerns the effect of overflow afferent signals on a deafferented cortex. Ramachandran maintains that the effect of overflow afferent signals is the restructuring of a deafferented area, such that its structure becomes continuous with a neighboring intact cortical area. The preservation hypothesis proposes that overflow afferent stimuli contribute to the structural and functional preservation of a deafferented cortex.

In contrast to the remapping hypothesis, the preservation hypothesis maintains that a primary function of the brain is the preservation of the brain's structures and functional capacities, even if this conflicts with accurately representing the condition of the body. Following deafferentation, the brain engages in no activity that would eliminate phantom sensations. On the contrary, activities in the brain would work to preserve the neural structures giving rise to phantom sensations, thus maintaining the structures and functional capacities of the brain regardless of whether the consequence is an inaccurate experience of the body.

V. Remapping vs. Preservation: An Empirical Comparison

The preservation hypothesis proposes that cortical co-activation alerts us to a mechanism by which the brain continues to supply a deafferented cortex with afferent stimulation, contributing to that cortex's long-term structural and functional preservation. This is in direct opposition to the remapping hypothesis, which proposes that the mechanism indicative of cortical co-activation simultaneously serves to deconstruct a deafferented cortex and reorganize it to be continuous with the structure and functional capacities of a neighboring intact cortex. Aside from the fact that afferent stimulation can arrive at a deafferented cortex from numerous other cortical areas with the effect of contributing to the long-term preservation of the deafferented cortex, is there any other empirical evidence that could support the preservation hypothesis? Indeed, there is.

A serious challenge facing the remapping hypothesis is the need to explain certain discrepancies between the entailments of the remapping hypothesis and the phenomenal experiences of amputees with phantom sensations. For instance, the remapping hypothesis posits that the cortical reorganization that follows deafferentation dismantles the neural structure that gives rise to phantom sensations, resulting in a significant reduction in the vividness of phantom sensations and, ultimately, the complete disappearance of phantom sensations. Conversely, we ought to expect an increase in the sensitivity and/or functional capacity of intact cortices that now have a greater area of representation in the brain.

Unfortunately for the remapping hypothesis, this progressive elimination of phantom sensations is just not experienced among amputees with phantom limbs. Amputees typically experience their phantom limbs long-term, without any decrease in the vividness of the phantom. A study by Lotze *et al.* on the effects of myoelectric prosthetics of the somatosensory cortex includes a patient who continues to experience phantom sensations 53 years after amputation (Lotze *et al.* 1999). Another study by Makin *et al*. includes a patient who continues to experience phantom sensations 47 years post-amputation, with the average post-amputation time of all 18 of their subjects being 18 years (Makin *et al*. 2012). And even after extended periods of time following amputation, amputees not only continue to feel their phantoms but can execute fine motor skills, such as the opening and closing of their phantom fists and "thumb to index opposition" (Raffin, E. *et al*. 2012, 748). Additionally, Ramachandran proposes that we ought to see functional enhancements associated with body parts whose area of representation has expanded into a deafferented area (Ramachandran 2000, 319). While more research may have to be conducted on this front, currently, no major research study

has found a noticeable increase in the functional capacity of intact body parts following deafferentation.

Phantom sensations do not disappear over time, do not decrease in vividness, and amputees tend to maintain motor control and the ability to execute fine motor skills even five decades, or more, after amputation. These experiences of phantom sensations do not sound like the perceptual correlates of dramatic cortical restructuring. Rather, these perceptual correlates appear perfectly consistent with continued structural and functional preservation of a deafferented cortex.

For Giummarra *et al*., the evidence overwhelmingly suggests that a deafferented cortex continues to receive afferent signals from diverse sources in the brain, and that the arrival of these afferent signals support the continued preservation of the deafferented cortex. However, Giummarra does not dispute the arguments that cortical co-activation points to "rapid cortical reorganization contralateral to the deafferented limb" (Giummarra *et al*. 2007, 227). The picture painted by Giummarra *et al*. is one of a brain divided. On the one hand, deafferentation initiates a process of neural restructuring, such that an intact cortex invades a neighboring deafferented area, utilizing pre-existing neural connections to expand its area of representation in the brain. On the other hand, a deafferented cortex continues to receive stimulation from diverse neural sources, with the effect of preserving some of its neural structure and functional capacities. The result is that the somatosensory cortex is subject to two simultaneous, yet contrary, mechanisms; one which erodes the structure of a deafferented cortex and another that preserves it. As Ramachandran, Giummarra *et al*. maintain that a primary function of the brain is the accurate representation of the body, but simultaneously suggest that the very structure and normal functioning of the brain itself stands in the way of the brain fulfilling that function.

Not all researchers, however, take cortical reorganization as a given. Tamar Makin, for instance, argues that dramatic cortical reorganization is not the effect of deafferentation, and that the evidence in support of it "is largely based on…crude measurements" (Makin *et al.* 2015, 2140). To remedy the often "crude" and inconsistent measurements of cortical co-activation,

> …we assessed remapping of sensorimotor lip representations using an unfolded model of the cortex, allowing us to measure surface-based cortical distances while considering individual cortical folding patterns (Maeda et al., 2014) in 17 unilateral upper limb amputees and 21 intact controls. We found consistent shifts in lip representation

> along the homunculus contralateral to the missing hand in amputees (hereafter 'deprived homunculus') towards the hand area. However, this shift didn't reflect full invasion of the lips into the hand territory as previously described, but rather a small local shift in the centre of gravity of the lips (Makin *et al.* 2015, 2141).

Using an "unfolded model of the cortex," Makin *et al.* determine that there is a shift in contralateral intact lip representation in unilateral hand amputees, but that this shift is much smaller than previously documented and does not constitute a "full invasion of the lips into" the deafferented cortex. Measured "shifts in lip representation" were consistent among unilateral hand amputees, but so slight that Makin *et al.* could not establish "any statistical relationship between cortical reorganization and phantom sensations" (Makin *et al.* 2015, 2145). Makin *et al.*'s findings, while they certainly require further corroboration, open the possibility that cortical co-activation may not be as dramatic as previously thought. And, while Makin *et al.* do not posit why shifts in lip representation are observed, it is possible that these consistent, yet slight, shifts of lip representation toward the deafferented cortex are the effect of overflow afferent signals from the lip area successfully reaching and exciting the deafferented hand cortex by way of new or pre-existing neural connections, as suggested by the preservation hypothesis.

In addition to raising concerns over the extent of cortical reorganization/co-activation that is being reported in phantom limb studies, Makin is one of a handful of researchers spearheading a new approach in the way that phantom sensations are studied. As mentioned above, phantom sensation research has historically been challenging because its subject concerns a non-existing limb; the study lacks a peripheral limb at which to focus its inquiry. When Pons discovered that touching an intact body part could excite activity in a deafferented cortex, the field of phantom limb research exploded; finally, researchers could utilize the intact peripheral body as a means of studying sensations corresponding to a non-existent limb. In recent years, a new revolution in the study of phantom limb research has begun through the identification of an alternative medium by which to study phantom sensations. This medium; the movement of a phantom limb.

In their research, Karen T. Reilly and Estelle Raffin have sought to determine whether executive movements of a phantom limb are purely imaginary or resemble the executive movements of intact limbs. Reilly *et al.* and Raffin *et al*. measured EMG activity in the stumps of unilateral hand amputees and two-handed subjects both when they moved their phantom limbs and when they were asked to imagine moving their phantom limbs (Reilly *et al*. 2006; Raffin *et al.* 2012). No significant activity was recorded for either

amputees or two-handed subjects when asked to *imagine* moving their limbs. However, both Reilly *et al*. and Raffin *et al*. observed "significant movement-related bursts of EMG activity in stump muscles" when unilateral hand amputees completed executive motions with their phantom limb (Raffin *et al*. 2012, 753). Additionally, amputees "insist" that imagined movements of their phantom limbs feel like imagined movements of other intact body parts, whereas "motor execution with the phantom evokes sensations close to those experienced when they actually move a body part" (Raffin et al, 754-5). Raffin *et al*. conclude that "amputees moved their phantom limb during our execution condition and imagined moving it during our imagination condition" (Raffin et al, 753).

The possibility of observing the effects of deafferentation in the brain by way of patients "moving" their phantom limbs is being pursued further by Tamar Makin. Using 18 unilateral upper-limb amputees with an average of 18 years since amputation and "22 intact controls (two handers)," Makin *et al*. conducted a series of fMRI scans to determine neural activity that corresponded exclusively to the movement of a phantom limb (Makin *et al*. 2012, 2). Makin *et al*. found that "group activation for phantom movements was similar to that found during two-handers' non-dominant hand movements in the primary sensorimotor cortex…suggesting preserved functional representations" (Makin *et al*. 2012, 2-3). Prior to Makin *et al*.'s research, we have never observed neural activity isolated to the deafferented cortex; we could only estimate the effects of deafferentation by way of the phenomenon of cortical co-activation. Makin *et al*.'s focus on phantom movements themselves, has finally provided a means to isolate neural activity in the deafferented cortex, and the findings are such that there is a great degree of preservation of the cortex's original area of representation.

Makin *et al*.'s findings certainly prove problematic for the remapping hypothesis. If cortical co-activation directly corresponded with cortical reorganization, then neural activity in the deafferented cortex ought to be significantly reduced in size compared to the size of the cortical territory prior to deafferentation. However, that's just not what we see; when we observe neural activity isolated to the deafferented cortex, we observe preservation of cortical structure. Now perhaps something like what Giummarra *et al*. have suggested is occurring, that is, following deafferentation, there are two mechanisms at work simultaneously, one which deconstructs a deafferented cortex and another that preserves the structure of a deafferented cortex. This may very well be the case, and if it were true could partially vindicate the remapping hypothesis. Something like remapping may be occurring, it just would take much longer to obtain because of counteracting forces. But, at the very least, the remapping hypothesis and researchers must concede a critical piece of evidence on which the remapping hypothesis rests; cortical co-activation

does not directly correspond to cortical reorganization. The remapping hypothesis takes cortical co-activation to be identical with cortical reorganization, such that cortical co-activation indicates the new boundaries of an intact cortical area and the extent to which it has invaded a deafferented cortex. However, if neural activity isolated to the deafferented cortex shows a preservation of the original neural structure, then clearly cortical co-activation does not align with cortical reorganization.

Makin *et al.*'s findings could also be indicative of another possibility, that something like what the preservation hypothesis proposes is at work following deafferentation. Note that the preservation hypothesis argues that cortical co-activation is not indicative of cortical reorganization but reveals intercortical transference of afferent stimuli with the effect of preserving the structure and functional capacities of a deafferented area. Thereby, if we were ever able to observe the isolated activity of a deafferented cortex, we ought to see a great degree of structural preservation. And that, in fact, is exactly what we observe, now that Makin *et al.* have pioneered a means to isolate the neural activity of a deafferented cortex.

VI. Concluding Remarks

My intent in writing this piece is not to indisputably prove the preservation hypothesis or debunk the remapping hypothesis. What I want to show is that there is sufficient evidence to suggest that something like the preservation hypothesis could explain research findings on the effects of deafferentation in the somatosensory cortex. And, if this is the case, then additional research ought to pursued that explores the possibility of neural preservation following deafferentation.

Whether remapping or the preservation of a cortex follows deafferentation is of material consequence to the determination of how best to treat phantom limb pain. Researchers and clinicians who take remapping seriously, tend to approach the medical treatment of phantom pain in terms of (1) expediting cortical reorganization, and/or (2) the pharmaceutical management of pain.[13] The preservation hypothesis conceives of the health of the brain in terms of its ability to preserve all its constituent structures and functional capacities, suggesting that treatments for phantom pain ought to be directed towards accomplishing that end.

13. Medications currently used to treat phantom pain include "opioids, NMDA receptor antagonists, anticonvulsants, antidepressants, calcitonins, and anaesthetics," with researchers finding these pharmaceutical interventions "unsatisfactory" in managing phantom pain long-term (Alviar *et al.* 2011, 2).

Additionally, phantom sensation research provides critical insights for conceiving a primary function of the brain: does the functioning of the brain produce an accurate representation of the body or preserve the structures and functional capacities of the brain/body? The phenomenon of phantom sensation pits these two accounts of brain function against one another. Eliminating phantom sensations and restoring the alignment between the experience and actual condition of the body requires the destruction of the neural structures that underpin the appearance of a phantom limb. Thereby, the process of ensuring the accurate representation of the body would come at the cost of the preservation of the brain's structures and functional capacities. In contrast, the brain's operating to preserve its neural structures and functional capacities, the very neural structures that give rise to phantom sensations, would result in the continued discordance between the experience and actual condition of the body. Given that the requirements for accurate representation and preservation each entail a condition that would prevent the other from obtaining, it would be impossible for the brain to simultaneously pursue both these ends. If the brain's function is to preserve itself, one's experience of the body will be in permanent discord with the actual condition of the body. And if the brain's function is to accurately represent the body, it must engage in self-destruction, eliminating the neural structures of the phantom limb to restore the alignment between our experience of the body and the body's actual condition.

It is important to note that these philosophical musings over the function of the brain are very much of practical significance. As Socrates points out in the *Phaedrus*, if we want to make the body "healthy and strong" on "the basis of an art," then it is necessary "to determine the nature of…the body" (Plato *Phaedrus*, 270b3-7). In the case of neuroscience, we must extend Socrates' insights to include the brain as well, such that, in pursuing an understanding of the nature of the brain, we simultaneously pursue an understanding of what truly constitutes a healthy and strong brain. Thus, I do hope that my work here encourages future research projects on phantom sensations. But even more so, I hope that this piece inspires researchers, in all areas of neuroscience, to reflect upon their conceptual commitments concerning the ultimate function of the brain, and to consider how these commitments affect their research and approach to the treatment of neurological conditions.

References

Alviar, MJM., Hale, T., and M. Duncan. 2011. "Pharmacological interventions for treating phantom limb pain." Cochrane Database of Systematic Reviews 12 (CD006380): 1-54.

Descartes. 1998. *Discourse on Method* and *Meditations on First Philosophy*. Translated by Donald A Cress. Indianapolis: Hackett Publishing Company.

Descartes. 1998. *The World and Other Writings*. Translated and Edited by Stephen Gaukroger. Cambridge: Cambridge University Press.

Finger, S. and Hustwit, M. 2003. "Five Early Accounts of Phantom Limb in Context: Pare, Descartes, Lemos, Bell, and Mitchell." *Neurosurgery* 52 (3): 675-686.

Giummarra, Melita J., Gibson, Stephen J., Georgiou-Karistianis, Nellie, and John L. Bradshaw. 2007. *Brain Research Reviews* 54: 219-232.

Hassing, Richard. 2015. *Cartesian Psychophysics and the Whole Nature of Man*. Lanham: Lexington Books.

Lotze, M., Grodd, W., Birbaumer, N., Erb, M., Huse, E., and H. Flor. 1999. "Does use of a myoelectric prosthesis prevent cortical reorganization and phantom limb pain?" *Nature Neuroscience* 2 (6): 501-502.

Makin, Tamar R., Scholz, Jan, Filippini, Nicola, Henderson Slater, David, Tracey, Irene, and Heidi Johansen-Berg. 2013. "Phantom pain is associated with preserved structure and function in the former hand area." *Nature Communications* 4 (1570): 1-8.

Plato. 2006. *Plato on* Love. Edited by C. D. C. Reeve. Indianapolis: Hacking Publishing Company, Inc.

Pons, TP., Garraghty, PE., Ommaya, AK., Kaas, JH., Taub, E., and M. Mishkin. 1991. "Massive cortical reorganization after sensory deafferentation in adult macaques." *Science* 25 (5014): 1857-60.

Raffin, Estelle, Giraux, Pascal, and Karen T. Reilly. 2012. "The moving phantom: Motor execution or motor imagery?" *Cortex* 48: 746-757.

Reilly, KT., Mercier, C., Schieber, MH., and A. Sirigu. 2006. "Persistent hand motor commands in amputees' brain." *Brain* 129 (8): 2211–23.

Ramachandran, V.S. 1998. *Phantoms in the Brain*. New York: Quill.

Ramachandran, V.S. 2011. *The Tell-Tale Brain.* New York: W. W. Norton & Company.

Ramachandran, V.S. and W. Hirstein. 1998. "The perception of phantom limbs: the D.O. Hebb lecture." *Brain* 121: 1603-1630.

Ramachandran V.S., Rogers-Ramachandran, D, and M. Stewart. 1992. "Perceptual correlates of massive cortical reorganization." *Science* 258 (5085): 1159–60.

Ramachandran, V.S., and D Rogers-Ramachandran. 2000. "Phantom Limbs and Neural Plasticity." *Arch Neurol.* 57 (3): 317-320.

Sherman, RA., Sherman, CJ., and L. Parker. 1984. "Chronic phantom and stump pain among American veterans: Results of a survey." *Pain* 18 (1): 83–95.

Journal of Cognition and Neuroethics

How to Defend Embodied Cognition Against the Locked-In Syndrome Challenge

Luis H. Favela
University of Central Florida

Biography

Luis H. Favela is Assistant Professor of Philosophy and Cognitive Sciences at the University of Central Florida. His research is both philosophical and empirical, residing at the intersection of the cognitive sciences, neuroscience, philosophy, and psychology. His primary research aim is to demonstrate the suitability of complexity science and dynamical systems theory to provide the appropriate theories and methods for investigating and understanding mind, where 'mind' includes behavior, cognition, and consciousness.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). February, 2018. Volume 5, Issue 1.

Citation

Favela, Luis H. 2018. "How to Defend Embodied Cognition Against the Locked-In Syndrome Challenge." *Journal of Cognition and Neuroethics* 5 (1): 27–48.

How to Defend Embodied Cognition Against the Locked-In Syndrome Challenge

Luis H. Favela

Abstract

Embodied cognition is the idea that cognition is causally related to and/or constituted by bodily activities. In spite of accumulating reasons to accept embodied cognition, critics seem to have a knockdown argument: appealing to locked-in syndrome (LIS). Patients with LIS are said to be at least minimally conscious to fully awake, except they have no motor control of their body and cannot produce speech. LIS seems to undermine embodied cognition: if cognition is embodied, then LIS patients cannot have intact cognitive capacities because they do not have motor control of their body. The present goal is to provide supporters of embodied cognition with a set of three responses when faced with the challenge from LIS. The first is deflationary and highlights the fact that most cases of LIS are not total and that much evidence of LIS are not actually cases of LIS. The second is skeptical and provides reasons to question the evidence of LIS based on neuroimaging data. The third is that the types of pathologies that cause LIS are likely to alter cognition in radical ways. With these responses at the ready, the supporter of embodied cognition need not surrender at the mere mention of LIS.

Keywords

Cognition, Consciousness, Embodied Cognition, Locked-In Syndrome

Introduction

If there is a generally accepted understanding of cognition, then it is that cognition is a phenomenon that happens in brains and is essentially information processing and representational in nature (Kandel, Schwartz, Jessell, Siegelbaum, and Hudspeth 2013). For the past few decades, this "received view" of cognition has experienced some pushback. The term 'cognition' has come to be applied in non-brain-centric ways, for example, "embodied cognition," or the idea that cognition is causally related to and/or constituted by non-neuronal physiology (Varela, Thompson, and Rosch 1991). One reason given in support of embodiment is that cognition involves bodily action. Be it for purposes of cognitive development (Thelen and Smith 1994) or simulating the mental states of others (Barsalou 2008), proponents of various forms of embodiment argue that cognition relies on bodily states and movements to varying degrees. Such non-brain-centric approaches to cognition are receiving increasing theoretical support in philosophy and empirical evidence in the cognitive, neural, and psychological sciences.

In spite of the evidence and reasons to accept non-brain-centric conceptions of cognition, critics seem to always have a knockdown argument, especially against embodied cognition. The argument centers on appealing to cases of locked-in syndrome (LIS). Patients with LIS are said to be at least minimally conscious to fully awake, except they have no motor control of their body and cannot produce speech. LIS seems to undermine embodied cognition in this way: If cognition is embodied, then LIS patients cannot have intact cognitive capacities because they do not have motor control of their body. The present goal is to provide supporters of embodied cognition with a set of three responses when faced with the ever-looming challenge from LIS. The first response is deflationary and highlights the fact that most cases of LIS are not total and that much evidence of LIS are not actually cases of LIS. The second is skeptical and provides reasons to question the evidence of LIS based on neuroimaging data. The third is that the types of pathologies that cause LIS are likely to alter cognition in radical ways.

In the next section, I present an overview of the received view of cognition and alternative understandings, with a focus on embodied cognition. Next, I present LIS and explain why it is a challenge for proponents of embodied cognition. After, I present responses to the challenge of LIS. With these responses at the ready, the supporter of even the more radical versions of embodied cognition need not surrender at the mere mention of LIS.

The Received View of Cognition and Some Alternatives

It is likely safe to say that most contemporary researchers of cognition in philosophy and the relevant sciences believe that cognition happens in brains (Kandel, Schwartz, Jessell, Siegelbaum, and Hudspeth 2013). Additionally, and at least since the "cognitive revolution" of the 1950s, most of those folks believe that cognition is essentially representational and involves information processing (Miller 2003; Thagard 2005). In regard to behavior, cognition is understood not as a kind of behavior, but as a cause of behavior (Aizawa 2015; Fodor 2008; 2009; Shapiro 2013). Taken together, these four characteristics comprise the contemporary "received view" of cognition: brain-centered, cause of behavior, information processing, and representational. Though popular in textbooks, accounts that present the received view as the only game in town concerning scientific investigations of cognition since the 1950s are false and incomplete. Other approaches to understanding cognition were concurrently in development and practiced alongside the received view.

There is no doubt that research programs investigating cognition based on computational-representational frameworks were heavily influential at least from the 1950s to 1980s (e.g., Good Old Fashioned Artificial Intelligence [GOFAI]), and that investigations of the brain have taken a more central role from the 1990s to today (e.g., connectionism, neural networks, and neuroimaging [Boden 2006]). However, concurrent with such methodologically solipsistic computational-representational approaches (Fodor 1980) were frameworks that shifted focus away from cognition isolated in brains to cognition in *systems* (Favela and Martin 2017). It has been claimed that conceptions of cognition as not isolated in individuals—or, specifically, their brains—has roots in Darwinian biology and Jamesian psychology of the late-1800s to early-1900s (Chemero 2009; Favela and Martin 2017). For current purposes, I focus on ecological psychology as a starting point for thinking about cognition as a systems phenomenon.

There are many available summaries of James Gibson's ecological psychology (e.g., Chemero 2009; Favela and Chemero 2016a; Richardson, Shockley, Fajen, Riley, and Turvey 2008). What matters for now is that Gibsonian ecological psychology, which began in the mid-1900s, provides a theoretically rich and empirically successful alternative to the received view, and one that influenced many of the current non-brain-centric alternatives. Ecological psychologists reject many of the central tenants of the received view. First, cognition, action, and perception are not treated as distinct but as continuous. Accordingly, the cognition/behavior dichotomy is understood as false. Second, cognition is not computational/information processing in nature, representational, or isolated in brains. For the ecological psychologist, cognition is understood as occurring across organism-environment systems. Third, cognition-action-perception are temporal in nature. One favorite example of the ecological psychologist that demonstrates these three features is the outfielder problem.

In short, the outfielder problem is the challenge of explaining how a baseball player can catch a ball that is moving high in the sky. Based on its theoretical commitments, the received view would have to say that the player creates a mental representation of the ball in the sky, computes her location relative to the ball, then—based on stored information cuing features of the environment (e.g., depth based on surface orientation [Marr (1982) 2010])—moves her body towards the ball, updates her mental representations of the ball, recalculates the position of the ball relative to her body, and so on, until her neurally-realized calculations bring her gloved hand to the ball. Alternatively, the ecological psychologist posits a much more parsimonious explanation: the player directly perceives the ball in the sky, and based simply on the changing size of the ball in the sky—that is, the ball appears larger as it gets closer and smaller when further—

moves in its direction and to catch it. As ecological psychologists have argued (Oudejans, Michaels, Bakker, and Dolne 1996), such accounts that take environmental information into consideration in that way provide non-brain-centric, non-computational, and non-representational explanations. Instead of computing indirect mental representations, the ecological explanation offloads information onto the environment (i.e., object size occlusion) and utilizes temporal features (e.g., parallax). The emphasis on the temporal dimension of cognition-action-perception facilitated a fruitful combining of ecological psychology with dynamical systems theory.

Dynamical systems theory (DST) originated with Newton's invention of differential equations to help explain planetary motion. The emphasis placed on the temporal properties of phenomena lead some in the cognitive sciences to wonder whether cognition is dynamic in nature or, at the very least, whether the methods of DST could illuminate understanding of cognition (e.g., van Gelder 1998). The typical DST treatment of a phenomenon includes capturing the relevant variables within differential equations and plotting the phenomenon in a phase space, which represents all possible states of the system over time. When thinking about DST approaches, it helps to keep in mind their emphasis on principles of behavior. Instead of decomposing a phenomenon to see what it is made of and what is the primary/first cause of its behavior, DST approaches focus on how the states of the system evolve over time according to a rule (Riley and Holden 2012). These rules (or principles) are often written with differential equations, which include order parameters (macroscopic states of system), and control parameters (variables guiding system dynamics [Haken (1988) 2006]). Two features of DST explanations are especially important for current purposes. First, from an explanatory perspective, DST models do not make a priori discriminations about where in the world things represented by variables should be located. Second, and following from the first, since DST accounts do not dictate where boundaries ought to be, what counts as a "system" can sometimes be counterintuitive (Beer 1995; Chemero 2009).

Consider, for example, the following coupled differential equations:

(1) $\dot{x} = x + 2y$

(2) $\dot{y} = 3x + 2y$

These two equations are "coupled" in the sense that any changes to one will affect the other. Thus, if the value of x in (1) is increased and thereby resulting in an increase in $\dot{x}$, then so too will there be an increase in x in (2), which will thereby result in an increase in $\dot{y}$. Suppose that (1) and (2) are two coupled equations that model and explain how one

person's arms moves while clapping. It would be easy to say the equations depict a single system—(1) represents the left arm and (2) the right arm. Now, suppose that (1) and (2) are two coupled equations that model and explain how one person's arm (x) moves while tossing a ball (y) in the air and catching it. That might be more difficult to accept as a "single system" for various reasons, for example, because the ball is inorganic, not attached to the person, etc. Yet, as the model indicates, any changes to one variable (y) is determined by changes in the other (x). I do not know if dynamicists would typically accept that the ball-and-arm count as one system. However, the example is instructive in its ability to flesh out what DST has pushed investigators of cognition to consider, especially the location of factors relevant and constitutive of cognition.

Given the explanatory virtues DST approaches facilitate (e.g., controlled manipulations, predictions, etc.), it is not so easy to dismiss their sometimes-counterintuitive consequences. For example, given the nature of the relationships among variables depicted by DST models and plots, it may be necessary to rethink whether a phenomenon is merely a set of distinct but coupled components, or if those components are so tightly related that they *constitute* the phenomenon. It is this ability to model systems with sometimes-counterintuitive variables that mesh so well with ecological psychology (Favela and Chemero 2016b; Kugler, Kelso, and Turvey 1980). As mentioned above, ecological psychologists explain events such as fly-ball catching with an eye towards the system, and not isolated organisms. As a result, the baseball player, field, and ball play roles in complete solutions to the outfielder problem. With the tools of DST, the ecological psychologist can explain how such variables can (even counterintuitively) constitute one system, which can be evidenced by such empirical findings as alterations to one part of the system (e.g., ball location) affecting other parts of the system (e.g., player location). The theoretical and methodological successes of ecological psychology and DST (taken together and alone) prompted a range of reasons to think that cognition is not just in the head.

Along with the received view, fruitful alternative frameworks were conducting research guided by such theories as those found in ecological psychology and methods such as DST. One consequence of such frameworks has been the possibility that cognition is not centered in the brain, nor computational or representational. A number of non-brain-centric approaches began to surface in the 1980s and 1990s. These included treatments of cognition as distributed, embodied, enactive, extended, and situated. Such views have not merely been argumentative or theoretical considerations, but have become increasingly influential in philosophy and the cognitive, neural, and psychological sciences (for reviews see Favela 2014; Favela and Chemero 2016b; Favela and Martin

2017). Here I focus on embodied cognition. Embodied cognition is a non-brain-centric position concerning what causes and constitutes cognition (Anderson 2003; Richardson, Shockley, Fajen, Riley, and Turvey 2008; Rowlands 2010; Varela, Thompson, and Rosch 1991; Wilson 2002). It is important to note that "embodied cognition" is not equivalent to "embodied mind" (Favela and Chemero 2016b). Embodied mind is a metaphysical thesis about the nature of mental states. For example, the idea that mental states occur not just in the brain but in the body as well can be viewed as functionalist in nature, namely, that mental states are defined by particular realization relationships that extend into the body (cf. Wilson 1994). Embodied cognition, on the other hand, does not necessarily make claims about the metaphysics of mind, and could be consistent with various metaphysics such as eliminativism, functionalism, or identity theory.

Although there is no single "embodied cognition," the general thesis is that the body's sensory and motor processes constrain and enable cognition (Foglia and Wilson 2013; Thompson 2007). There is a wide range of consequences that thesis has for how we consider cognition. Conservative versions of embodied cognition treat cognition as computational and representational in nature (Barsalou 2008; 2010; Wilson 1994). These conservative versions treat cognition as essentially representational in nature, but the representations are not necessarily realized neuronally or of a single kind, for example, the body can temporally represent states such as another's pain (Barsalou 2008). Others claim that cognition can extend to tools and the environment (Clark and Chalmers 1998; Fiore and Wiltshire 2016; Hutchins 1995). Such extended and distributed forms of cognition can be conservative and remain consistent with cognition as computational-representational. For example, addresses in one's smartphone can serve as external representations that can be accessed via information-processing means (i.e., functionalism [Clark and Chalmers 1998]). Finally, there are radical versions that reject the idea that cognition involves any information-processing or representations. Instead, cognition is non-computational and non-representational, and is fundamentally constituted by the dynamics of brain-body-environment systems (Chemero 2009; Kelso 1997; 2009; Port and van Gelder 1995; Thelen and Smith 1994). Whether one adheres to radical conceptions of embodiment or not, embodied cognition has become an influential branch of cognitive science (Calvo and Gomila 2008; Chemero 2009; Dale 2008; Favela and Martin 2017; Glenberg 2010; Riley, Shockley, and Van Orden 2012). What is more, even researchers in the neural sciences are acknowledging both the causal and constitutive roles the body plays regarding cognition (Edelman 2006; Favela 2014; Sporns 2010; Tognoli and Kelso 2014).

Until now, I have attempted to motivate the case that various substantial research programs have investigated cognition as long as the received view has. Some of these programs share features of the received view—for example, that cognition involves some form of information processing—but some are radically different. For example, ecological psychology rejects the received view's treatment of cognition, action, and perception as distinct, and DST facilitates understanding cognition as fundamentally temporal in nature. Taken together, ecological psychology and DST have served as precursors for non-brain-centric conceptions of cognition. Various theories and methods now investigate cognition as, for example, distributed, extended, and situated. One of the more robust forms of non-brain-centric cognition is embodied cognition, which continues to gain increasing theoretical and empirical support across philosophy and the mind sciences. Still, there are many critics of embodied cognition, especially of the radical sort that rejects understanding cognition as computational or representational. One knockdown argument against embodied cognition—especially radical versions—at the critic's disposal is the appeal to locked-in syndrome, which I turn to in the next section.

The Locked-In Syndrome Challenge

In general, a patient has locked-in syndrome (LIS) when she is fully awake but cannot move her body or verbally communicate. The LIS challenge to embodied cognition can be understood as primarily motivated by the following question: if cognition is embodied, then how can patients with LIS have intact cognitive capacities despite having no motor control of their body? The argument is as follows: First, embodied cognition claims that cognition is causally related to and/or constituted by the body's sensorimotor activity. Second, LIS patients have intact cognition despite being unable to move their body. Third, LIS patients have intact cognition without their cognition being embodied. Therefore, cognition is not embodied. Thus, embodied cognition is an incorrect theory about the nature of cognition. In the remainder of this section, I will explain in more detail what LIS is and why it is more of a challenge for some forms of embodied cognition than others.

Patients with LIS are at least minimally conscious to fully awake, except they have no motor control of their body and cannot produce speech (Owen 2013). LIS typically results from strokes (86.4%; [León-Carrión, Van Eeckhout, Dominguez-Morales, and Perez-Santamaria 2002]), and is caused by:

> a primary vascular or traumatic injury to the brainstem, normally corresponding to a ventral pons lesion due to an obstruction of the basilar artery, and characterized by upper motor neuron quadriplegia,

> paralysis of lower cranial nerves, bilateral paresis of horizontal gaze and anarthria, and with preserved consciousness. (León-Carrión, Van Eeckhout, Dominguez-Morales, and Perez-Santamaria 2002, 571)

LIS is not considered a disorder of cognition (Schnakers, Majerus, Goldman, Boly, Van Eeckhout, Gay, Pellas, et al. 2008), that is, LIS patients do not exhibit cognitive deficits such as impaired intelligence or memory. Additionally, LIS is not considered a disorder of consciousness (Owen 2013), that is, LIS patients are able to be awake and to distinguish sleeping from waking states. Though patients with LIS cannot produce speech, many can produce sound (78%; León-Carrión, Van Eeckhout, Dominguez-Morales, and Perez-Santamaria 2002) and the majority have vertical eye movements (Schnakers, Laureys, and Boly 2013), which means they have nonverbal communicative abilities. However, due to complete immobility that includes eye movement, some patients are diagnosed with *total LIS* (TLIS; Bauer, Gerstenbrand, and Rumpl [1979]).

TLIS is extremely difficult to diagnose. To be LIS is to not exhibit deficits in cognition or consciousness. This can be relatively straightforward to diagnose, as most LIS patients can communicate nonverbally, either by sound or eye movements. Thus, rudimentary forms of communication can be utilized to assess states of consciousness and cognitive capacities, for example, by moving eyes in a particular direction a certain number of times to indicate a letter in the alphabet. However, TLIS cannot utilize even those rudimentary means. Thus, it can be very challenging to diagnose a patient with TLIS as opposed to persistent vegetative state, which is a disorder of consciousness (The Multi-Society Task Force on PVS 1994). One approach to diagnosing TLIS in non-communicative patients is via functional magnetic resonance imaging (fMRI) to assess neuronal responses while the patient listens to spoken sentences (Owen, Coleman, Boly, Davis, Laureys, and Pickard 2006). In such experiments, speech-specific activation is assessed in areas of the brain that activate when non-TLIS subjects hear similar sentences. However, as Owens and colleagues (2006) point out, just hearing sounds and having accompanied neural activation does not mean the subject has conscious awareness of the sentences, that is, the subject's brain could merely have nonconscious sound or semantic processing. Accordingly, more sophisticated tests are needed, such as asking a potential TLIS patient to conduct mental imagery tasks in order to modulate their own neural activity in a manner that may not be as likely to result from automatic and/or nonconscious processing (Owen, Coleman, Boly, Davis, Laureys, and Pickard 2006). If such tasks are successful, namely, if TLIS patients can communicate mental states—for

example, fMRI detection of activity in neural areas following responses to task prompts such as "imagine playing tennis"—then it could be a major blow to embodied cognition.

Such findings would undermine embodied cognition because it would suggest that cognition is not sufficiently caused or constituted by the body, let alone necessarily so. Conservative forms of embodied cognition may be more readily poised to respond to the TLIS challenge. Perhaps the body was necessary in the development of cognitive capacities (cf. Thelen and Smith 1994) such as simulating states of others (cf. Barsalou 2008), but, once that ability is acquired, then the capacity can occur offline, that is, without the body. On the other hand, TLIS appears to be particularly devastating to radical embodied cognition, for it is committed to the idea that cognition is necessarily bodily: no sensorimotor capacities means no cognition. It appears that the thesis of radical embodiment—namely, that cognition is necessarily sensorimotor in nature—is disproven by TLIS patients who can conduct cognitive tasks offline by thinking about it in their head and without any body movement. Is it time then for proponents of radical embodied cognition to throw in the towel? No, proponents of radical embodied cognition are not doomed by the LIS challenge. In the next section, I provide three kinds of responses that proponents can offer when faced by the LIS challenge.

Saving Embodied Cognition from the Locked-In Syndrome Challenge

In the previous section, I presented the locked-in syndrome (LIS) challenge to embodied cognition: if LIS patients have intact cognitive capacities, then embodied cognition cannot be a correct theory of cognition. The apparent evidence of intact cognitive capacities in total LIS (TLIS) patients appears to make matters even worse for proponents of radical embodied cognition. In this section, I provide supporters of embodied cognition with a set of three responses to the LIS challenge: the first response is deflationary; the second is skeptical; and, the third raises concerns about the equivalence of cognitive states had by patients with TLIS compared to those without.

The first response to the LIS challenge is deflationary; specifically, most cases of LIS are not total and key cases in support of TLIS are actually not cases of LIS. The majority of patients with LIS have some degree of body movement (e.g., eye movement [Schnakers, Laureys, and Boly 2013]) and can communicate. In fact, many cases appealed to in the LIS challenge are not LIS at all. Take the "imagine playing tennis" example from Owen and colleagues' research (2006). As mentioned above, in order to attempt to control for detecting only nonconscious processing, Owen and colleagues asked potential TLIS patients to intentionally participate in tasks involving mental imagery. For example,

if a patient was known to enjoy playing tennis, then she was asked in experimentally controlled ways to imagine playing tennis, and if via fMRI scans neural activation was detected in motor areas of the brain, then it was presumed to be evidence that the patient was consciously aware enough to conduct that cognitive task. However, that particular set of experiments by Owen and colleagues was not intended to be a test for TLIS. It was a test, not of cognitive states, but of consciousness in patients in *persistent vegetative states* (PVS). Unlike T/LIS, which is categorized not as a disorder of cognition or consciousness, PVS is a disorder of consciousness: "The term describes a unique disorder in which patients who emerge from coma appear to be awake but show no signs of awareness" (Owen, Coleman, Boly, Davis, Laureys, and Pickard 2006, 1402). Moreover, PVS is a "clinical condition of complete [conscious] unawareness of the self and the environment, accompanied by sleep-wake cycles" (The Multi-Society Task Force on PVS 1994, 1499). Other disorders of consciousness that are not cases of T/LIS include comas, minimally conscious states, and brain death (Schnakers, Laureys, and Boly 2013). Thus, proponents of embodied cognition ought not to be swayed by many cases presented as evidence that LIS undermines embodiment because many of those cases are not actually LIS.

Another reason to deflate the significance of the LIS challenge is that such cases do not actually undermine the embodiment thesis in two key ways. Remember, embodied cognition is centered on the claim that cognition is causally related to and/or constituted by sensorimotor activity of the body. Moreover, such sensorimotor activity is necessarily temporal—see the above discussion of ecological psychology and dynamical systems theory. In many ways, patients with LIS still meet that criteria: "97.6% were temporally oriented" and "[n]early 100% of the patients reported being sensitive to touch to any part of their bodies" (León-Carrión, Van Eeckhout, Dominguez-Morales, and Perez-Santamaria 2002, 571). In short, nearly every LIS patient experiences their body in space and time, which is fundamental to the embodiment thesis. Thus, proponents of embodied cognition ought not to be so quick to equate a lack of motor control with an absence of bodily experience of the kind necessary to underlie cognition.

The second response to the LIS challenge is skeptical; specifically, there are good reasons to question how compelling the supposed evidence of TLIS is. Much of the evidence of TLIS relies on neuroimaging (e.g., Owen, Coleman, Boly, Davis, Laureys, and Pickard 2006; Pistoia and Sara 2012; Schnakers, Laureys, and Boly 2013). There are a number of deep methodological issues concerning neuroimaging that go far beyond the scope of the current work (e.g., Shulman 2013; Uttal 2001; 2011). In terms of defending embodied cognition against the LIS challenge, I focus on one major assumption involved

in interpreting neuroimaging data that may undermine such evidence as counting against the notion that cognition is embodied. That assumption is the presumed modular organization of the brain.

Neuroimaging experiments typically rely on the a priori assumption that mental states are modular and can be spatially and causally localized in the brain (Huettel and Song 2009). This claim assumes that if activity—such as blood flow in the case of fMRI—increases in an area of the brain during an experimental task, then that area of the brain is associated with a particular capacity. For example, if, during linguistic-related tasks, brain location X (e.g., Broca's area) exhibits increased blood flow, then that part of the brain is implicated with a linguistic capacity, such as language production. This assumption underlies a central method of the neural sciences: dissociations. Dissociations occur as follows: if location X is lesioned (e.g., a tumor in Broca's area is removed), and if a linguistic capacity is impaired (e.g., language production), then that is taken as even more evidence that location X is the primary location of those linguistic capacities. There are many practical reasons to justify the modularity assumption and method of dissociation (and double dissociations) in experimental practice. Attempting to decompose and localize parts of a system is often a first step in the scientific investigation of minimally understood and, often, highly complicated systems (Bechtel and Richardson [1993] 2010). The brain is one such minimally understood and highly complicated system. However, there are significant limits to the ability of such data from neuroimaging experiments that assume modularity to serve as evidence against embodied cognition.

The first limit I draw attention to concerns the circularity of such claims. As Van Orden and colleagues state:

> Modularity assumes morphological reductionism: Component effects reduce to underlying modules of mind and brain, and modules reduce to elementary causal microcomponents or *single causes*… Component effects are the structures of behavior, which are reduced to the structures of mind and brain. The assumption of single causes is the core assumption of modular research programs. (2001, 113; italics in original)

In other words, modularity assumes that mental phenomena have single causes, and when a particular phenomenon is made up of the combination of more basic capacities, then those more basic capacities have their own single causes. However, as Van Orden and colleagues forcefully argue, assuming modularity actually undermines the ability of dissociative methods to converge on fixed sets of exclusionary criteria to define

pure cases of dissociations (2001, 148). Such an inability of appealing to dissociations to locate modules ends up perpetually fractioning mental capacities into more and more modules and evermore finer grained locations (see Van Orden, Pennington, and Stone 2001 for detailed discussion). This is circular because built into the modularity assumption are theories about the nature of the mental states being dissociated. To claim that lesioning Broca's area dissociates language production capacities is to already have a theory about what language capacities are. Such theoretical commitments are not in themselves a weakness of neural sciences. Without definitions of concepts and theoretical commitments, an investigator would have no way of controlling an experiment or interpreting results. What makes modularity an unjustifiably circular assumption is that evidence of dissociations are searched for until they support a theory that is consistent with modularity. Consider the following example: Pierre Paul Broca had a patient named Louis Victor Leborgne who could vocally only produce the sound "tan" (Domanski 2013; Van Oden and Kloos 2003). If one is committed to modularity being the correct general theory about the structure of mental states in the brain, and if dissociations provide evidence for modularity, then Leborgne's ability to say "tan" while not being able to produce any other sound would be evidence for a "*tan* module" that is distinct from the other language module(s). If cases like Leborgne's were indeed evidence that dissociations bolster the case for mental modules, then modularity would be an absurd theory of mental states in light of an unjustified circularity of reasoning.

In addition to the problem of circular reasoning and its unintended absurd consequences (i.e., the "*tan* module"), the second limit of assuming modularity concerns the nonexistence of what I refer to as the "Cartesian module." Assuming mental modules can be localized in typically developed brains, there is a further issue concerning how brain lesions affect mental states. Perhaps typically developed brains share very similar modular organization at some compelling scale. If true, then the history of neuropsychology/science could very well be justified in appealing to dissociations to prove the existence of language modules, visual modules, reasoning modules, etc. However, the fact is that, after injury, the brain can reorganize so that mental capacities occur in sometimes very different gross anatomical areas of the brain—not to mention microscale areas. To a certain degree, even the proponent of modularity would agree that the brain reorganizes—after all, plasticity is necessary for learning. Nevertheless, what I point to now is more radical forms of neuronal degeneracy, which is the ability of structurally different elements to produce the same function or output (Edelman and Gally 2001). It becomes much more challenging to defend modularity when faced with the evidence of neuronal degeneracy, for example, numerical processing in varying

areas of the brains of different subjects (Krause, Lindemann, Toni, and Bekkering 2014), reorganized sensorimotor cortex in people born without particular limbs (Hahamy, Macdonald, Heiligenberg, Kieliba, Emir, Malach, Johansen-Berg, et al. 2017), and significant motor control without a cerebellum (Lemon and Edgley 2010). In addition to serving as considerations in opposition to the modularity theses, the previously stated examples of degeneracy also serve as evidence of the highly interconnected organization of the brain and, possibly, mental states.

What matters for current purposes is that the high degree of the brain's interconnectedness makes it very challenging to draw clear conclusions about the nature of the relationship among brain regions, cognitive ability, and conscious states (Bardin, Fins, Katz, Hersh, Heier, Tabelow, Dyke, et al. 2011). Consequently, and taken together with the first limit of modularity discussed above, it becomes unjustifiable to assume that whatever neuronal pathologies result in LIS have no effect on conscious thinking. For example, if modularity were true of neuronal capacities, then lesioning areas of the brain associated with motor control would have no effect on mental imagery related to bodily action. If true, then LIS patients would be in a "*Cartesian-like state*, in which thinking and acting are mutually dissociated" (Pistoia and Sara 2012, 2329; italics in original). In other words, for modularity to serve as a reason to disagree with the embodiment thesis, then it would be necessary for consciousness to reside in a "Cartesian module" where consciousness occurs in isolation. A Cartesian module would in this way be encapsulated from other brain and bodily properties, even those related to particular phenomenal states such as "visual processing modules" isolated from visual phenomenology that occurs in the Cartesian module. In summary, the second response to the LIS challenge is a set of reasons to be skeptical of the primary evidence of TLIS. That primary evidence is neuroimaging experiments that assume modularity, where modularity is circularly justified via dissociations. Additionally, the highly interconnected and degenerate nature of the brain's organization calls into question the ability to dissociate a "Cartesian module" of consciousness from modules related to particular kinds of phenomenal states.

Following from the second response, the third response to the LIS challenge raises concerns about the equivalence of cognitive states had by patients with TLIS compared to those who are not locked-in. To see why, we must first reject Cartesian consequences of modularity and dissociation-based evidence, specifically, that cognition and consciousness can remain unaltered even when dissociated from sensorimotor capacities. If cognition and consciousness cannot be dissociated from sensorimotor capacities without being altered themselves, then there should be significantly noticeable differences between

when they are and are not. If so, then we ought to consider the likelihood that patients with T/LIS have altered cognitive capacities and conscious experiences.

Remember, LIS is not categorized as a disorder of cognition or consciousness: a patient with LIS does not have any cognitive impairment and their conscious states are the same as when they were not locked-in. Therefore, the embodiment thesis is false. But then again, should we be so quick to accept that cognition and consciousness are unaltered in LIS? One set of reasons to not accept that cognition and consciousness are unaltered is a consequence of the second response to the LIS challenge given above. In short, the brain is so highly interconnected that it is unlikely that consciousness could exist in a Cartesian module that would be unaffected by alterations to other areas of the brain that it is connected to. Thus, if somebody has a stroke that results in damage to a part of their brain, then other areas would be affected as well due to the highly interconnected nature of the brain. If true, then a consequence of this claim is that damaging areas of the brain related to motor control would affect motor-control-related conscious experiences. To see why this is likely, consider cases of temporary motor-control paralysis.

Curare (*d*-Tubocurarine) is used by hunters to induce paralysis and, when combined with other anesthetics, to block pain during surgery. The paralyzed eye hallucination is an intriguing result that sometimes occurs due to curare use (Chemero and Cordeiro 2000; Favela and Chemero 2016a; Matin, Picoult, Stevens, Edwards, Jr., Young, and MacArthur 1982). Curare is used to induce temporary paralysis of voluntary movements in surgical patients. Upon waking after surgery, some patients are still incapable of voluntary motor-control that includes eye movement. Some of those patients report that when they tried to look around the room, their visual phenomenological experience was of the entire room moving in the direction they intended their eyes to move. For example, a patient's attempt to look left is unsuccessful because the muscles surrounding her eyes are paralyzed, though she has a visual experience of the whole visual field jumping to the left for a moment. One explanation of this phenomenological experience is that it is the result of perception being inextricably tied to embodied action. Though for a moment the patient has a visual experience of parts of her body, medical equipment, paintings on the wall, etc. jumping in the direction she intended to look, because there is no proprioceptive feedback, the visual experience is short-lived. Over time the patient's initial hallucinatory phenomenological experiences discontinue due to the lack of actual movement in the environment.

Accounting for the paralyzed eye hallucination scenario in this way serves as a response to the LIS challenge to embodied cognition. The primary reason is that it

provides a vivid example to support the claim that if patients with TLIS have cognition and consciousness, then it is likely that those capacities are altered in radical ways. By "altered in radical ways," I do not mean those states are deficient per se. I mean that the kind of cognition and consciousness had in locked-in states are unlike the kinds had when not locked-in. Thus, TLIS should not be considered a challenge to embodied cognition. The reason is that the embodiment thesis involves a set of claims (e.g., perception-action linked, cognition happens over time, etc.) about what is causally and constitutively relevant to cognition in particular systems. In cognitive systems like humans that are not locked-in, cognition and consciousness will be a particular way. This is the human perceptual life-world, or *umwelt* (von Uexküll [1934] 2010). As the paralyzed eye hallucination demonstrates, the umwelt of a human with TLIS will be radically different: the absence of motor-control results in unusual phenomenological experiences. Humans who do not have TLIS have kinds of visual experiences that do not include abrupt shifts of their entire field of vision like those experiences had when under the influence of curare. Consequently, due to the highly integrated nature of the brain, the injuries that caused TLIS have the consequence of altering the nature of locked-in patients' cognition and consciousness. This conclusion does not undermine the embodiment thesis, it provides more support for it.

Conclusion

The "received view" is that cognition is a brain-centered cause of behavior that involves information processing and is representational. Following work in ecological psychology and dynamical systems theory, the embodiment thesis claims that the body's sensory and motor processes constrain and enable cognition. Radical versions of embodied cognition claim that cognition does not involve any information processing nor representations, and is necessarily a systems-level phenomenon that involves brain, body, and environment interactions. In spite of evidence in support of the embodiment thesis, one objection seems to serve as a knockdown argument, especially against radical embodied cognition: locked-in syndrome (LIS). Patients have LIS when they are fully awake but have no voluntary motor-control and cannot verbally communicate, though they can move their eyes. Total LIS (TLIS) is when a patient cannot move their eyes either, thereby eliminating even rudimentary forms of communication made possible via eye movements. In short, the LIS challenge claims that cognition cannot be embodied (i.e., grounded in sensorimotor processes) because TLIS patients are presumed to have unimpaired cognition and consciousness despite having no motor control of their body.

I have attempted to provide three types of responses to the LIS challenge. Moreover, these responses are intended to save even radical embodied cognition from the challenge of TLIS. The first response is deflationary: most cases of LIS are not total and key cases in support of TLIS are actually not cases of LIS. The second response is skeptical: there are good reasons to question the ability of the current set of commitments of neuroimaging experiments to allow such data to serve as evidence of TLIS. The third response calls into question the equivalence of mental states had by patients with T/LIS and those that do not. This set of responses give proponents of embodied cognition—including the more radical varieties—reasons to not immediately raise a white flag in surrender at the mere mention of locked-in syndrome.

References

Aizawa, Ken. 2015. "What is this Cognition that is Supposed to be Embodied?" *Philosophical Psychology* 28: 755–775.

Anderson, Michael L. 2003. "Embodied Cognition: A Field Guide." *Artificial Intelligence* 149: 91–130.

Bardin, Jonathan C., Joseph J. Fins, Douglas I. Katz, Jennifer Hersh, Linda A. Heier, Karsten Tabelow, Jonathan P. Dyke, et al. 2011. "Dissociations between Behavioural and Functional Magnetic Resonance Imaging-Based Evaluations of Cognitive Function after Brain Injury." *Brain: A Journal of Neurology* 134: 769–782.

Barsalou, Lawrence W. 2008. "Grounded Cognition." *Annual Review of Psychology* 59: 617–645.

Barsalou, Lawrence W. 2010. "Grounded Cognition: Past, Present, and Future. *Trends in Cognitive* Science 2: 716–724.

Bauer, G., F. Gerstenbrand, and E. Rumpl. 1979. "Varieties of Locked-In Syndrome." *Journal of Neurology* 221: 77–91.

Bechtel, William, and Robert C. Richardson. (1993) 2010. *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. 2nd Edition. Cambridge, MA: MIT Press.

Beer, Randall D. 1995. "A Dynamical Systems Perspective on Agent-Environment Interactions." *Artificial Intelligence* 72: 173–215.

Boden, Margaret A. 2006. *Mind as Machine: A History of Cognitive Science*. New York, NY: Oxford University Press.

Calvo, Paco, and Toni Gomila. 2008. *Handbook of Cognitive Science: An Embodied Approach*. Amsterdam: Elsevier Science.

Chemero, Anthony. 2009. *Radical Embodied Cognitive Science*. Cambridge, MA: MIT Press.

Chemero, Anthony, and William Cordeiro. 2000. "Dynamical, Ecological Sub-Persons. Commentary on Susan Hurley's Consciousness in Action." In *A Field Guide to the Philosophy of Mind*, edited by M. Nani and M. Marraffa. Retrieved March 30, 2013 from http://host.uniroma3.it/progetti/kant/field/hurleysymp_chemero_cordeiro.htm

Clark, Andy, and David Chalmers. 1998. "The Extended Mind." *Analysis* 58: 7–19.

Dale, Rick. 2008. "The Possibility of a Pluralist Cognitive Science. *Journal of Experimental & Theoretical Artificial Intelligence* 20: 155–179.

Domanski, Cezary W. 2013. "Mysterious 'Monsieur Leborgne': The Mystery of the Famous Patient in the History of Neuropsychology is Explained." *Journal of the History of the Neurosciences* 22: 47–52.

Edelman, Gerald M. 2006. "The Embodiment of Mind." *Daedalus* Summer: 23–32.

Edelman, Gerald M., and Joseph A. Gally. 2001. "Degeneracy and Complexity in Biological Systems." *Proceedings of the National Academy of Science* 98: 13763–13768.

Favela, Luis H. 2014. "Radical Embodied Cognitive Neuroscience: Addressing 'Grand Challenges' of the Mind Sciences." *Frontiers in Human Neuroscience* 8 (796): 1–10. doi:10.3389/fnhum.2014.00796

Favela, Luis H., and Anthony Chemero. 2016a. "An Ecological Account of Visual 'Illusions.'" *Florida Philosophical Review* 16: 68–93.

Favela, Luis H., and Anthony Chemero. 2016b. "The Animal-Environment System." In *Foundations of Embodied Cognition: Volume 1: Perceptual and Emotional Embodiment*, edited by Y. Coelllo, and M. H. Fischer, 59–74. New York, NY: Routledge.

Favela, Luis H., and Jonathan Martin. 2017. "'Cognition' and Dynamical Cognitive Science." *Minds and Machines* 27: 331–355. doi:10.1007/s11023-016-9411-4

Fiore, Stephen M., and Travis J. Wiltshire. 2016. "Technology as Teammate: Examining the Role of External Cognition in Support of Team Cognitive Processes." *Frontiers in Psychology: Cognitive Science* 7 (1531). doi:10.3389/fpsyg.2016.01531

Fodor, Jerry A. 1980. "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology." *Behavioral and Brain Sciences* 3: 63–109.

Fodor, Jerry A. 2008. *LOT2: The Language of Thought Revisited*. Oxford: Oxford University Press.

Fodor, Jerry A. 2009. "Where is My Mind?" *London Review of Books* 31: 13–15.

Foglia, Lucia, and Robert A. Wilson. 2013. "Embodied Cognition." *WIREs Cognitive Science* 4: 319–325.

Glenberg, Arthur M. 2010. "Embodiment as a Unifying Perspective for Psychology." *WIREs Cognitive Science* 1: 586–596.

Haken, Hermann. (1988) 2006. *Information and Self-Organization: A Macroscopic Approach to Complex Systems.* 3rd Edition. New York, NY: Springer.

Hahamy, Avital, Scott N. Macdonald, Fiona van den Heiligenberg, Paullina Kieliba, Uzay Emir, Rafael Malach, Heidi Johansen-Berg, et al. 2017. "Representation of Multiple Body Parts in the Missing-Hand Territory of Congenital One-Handers." *Current Biology* 27: 1350–1355.

Huettel, Scott A., and Allen W. Song. 2009. *Functional Magnetic Resonance Imaging*. 2nd Edition. Sunderland, MA: Sinauer Associates, Inc.

Hutchins, Ewin. 1995. *Cognition in the Wild*. Cambridge, MA: MIT Press.

Kandel, Eric R., James H. Schwartz, Thomas M. Jessell, Steven A. Siegelbaum, and A. J. Hudspeth. 2013. *Principles of Neural Science*. 5th Edition. New York, NY: McGraw-Hill.

Kelso, J. A. Scott. 1997. *Dynamic Patterns: The Self-Organization of Brain and Behavior*. Cambridge, MA: MIT Press.

Kelso, J. A. Scott. 2009. "Coordination Dynamics." In *Encyclopedia of Complexity and Systems Sciences*, edited by R. A. Myers, 1537–1564. Berlin: Springer-Verlag.

Krause, Florian, Oliver Lindemann, Ivan Toni, and Harold Bekkering. 2014. "Different Brains Process Numbers Differently: Structural Bases of Individual Differences in Spatial and Nonspatial Number Representations." *Journal of Cognitive Neuroscience* 26: 768–776.

Kugler, Peter N., J. A. Scott Kelso, and Michael T. Turvey. 1980. "Coordinative Structures as Dissipative Structures I. Theoretical Lines of Convergence." In *Tutorials in Motor Behavior*, edited by G. E. Stelmach, and J. Requin, 3–70. Amsterdam: North Holland.

Lemon, R. N., and S. A. Edgley. 2010. "Life without a Cerebellum." *Brain: A Journal of Neurology* 133: 652–654.

León-Carrión, Jose, Philippe Van Eeckhout, Maria Del Rosario Dominguez-Morales, and Francisco Javier Perez-Santamaria. 2002. "Survey: The Locked-In Syndrome: A Syndrome Looking for a Therapy." *Brain Injury* 16: 571–582.

Marr, David. (1982) 2010. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge, MA: MIT Press.

Matin, Leonard, Evan Picoult, John K. Stevens, McIver W. Edwards, Jr., David Young, and Rodger MacArthur. 1982. "Oculoparalytic Illusion: Visual-Field Dependent Spatial Mislocalizations by Humans Partially Paralyzed with Curare." *Science* 216: 198–201.

Miller, George A. 2003. "The Cognitive Revolution: A Historical Perspective." *Trends in Cognitive Sciences* 7: 141–144.

The Multi-Society Task Force on PVS. 1994. "Medical Aspects of the Persistent Vegetative State (First of Two Parts)." *New England Journal of Medicine* 330: 1499–1508.

Oudejans, Raoul R. D., Claire F. Michaels, Frank C. Bakker, and Michel A. Dolne. 1996. "The Relevance of Action in Perceiving Affordances: Perception of Catchableness of Fly Balls." *Journal of Experimental Psychology: Human Perception and Performance* 22: 879–891.

Owen, Adrian M. 2013. "Detecting Consciousness: A Unique Role for Neuroimaging." *Annual Review of Psychology* 64: 109–133.

Owen, Adrian M., Martin R. Coleman, Melanie Boly, Matthew H. Davis, Steven Laureys, and John D. Pickard. 2006. "Detecting Awareness in the Vegetative State." *Science* 313: 1402.

Pistoia, Francesca, and Marco Sara. 2012. "Is There a Cartesian Renaissance of the Mind or Is It Time for a New Taxonomy for Low Responsive States?" *Journal of Neurotrauma* 29: 2328-2331.

Port, Robert F., and Timothy van Gelder. 1995. *Mind as Motion: Explorations in the Dynamics of Cognition*. Cambridge, MA: MIT Press.

Richardson, Michael J. Kevin Shockley, Brett R. Fajen, Michael A. Riley, and Michael T. Turvey. 2008. "Ecological Psychology: Six Principles for an Embodied-Embedded Approach to Behavior." In *Handbook of Cognitive Science: An Embodied Approach*, edited by P. Calvo, and T. Gomila, 161–187. Amsterdam: Elsevier Science.

Riley, Michael A., and John G. Holden. 2012. "Dynamics of Cognition." *WIREs Cognitive Science* 3: 593–606.

Riley, Michael A., Kevin Shockley, and Guy Van Orden. 2012. "Learning from the Body about the Mind." *Topics in Cognitive Science* 4: 21–34.

Rowlands, Mark. 2010. *The New Science of the Mind: From Extended Mind to Embodied Phenomenology*. Cambridge, MA: The MIT Press.

Schnakers, Caroline, Steven Laureys, and Melanie Boly. 2013. "Neuroimaging of Consciousness in the Vegetative and Minimally Conscious States." In *Neuroimaging of Consciousness*, edited by A. E. Cavanna, A. Nani, H. Blumenfeld, and S. Laureys, 117–131. Berlin Heidelberg: Springer-Verlag.

Schnakers, Caroline, Steve Majerus, Serge Goldman, Melanie Boly, Philippe Van Eeckhout, Stephane Gay, Frederic Pellas, et al. 2008. "Cognitive Function in the Locked-In Syndrome." *Journal of Neurology* 255: 323–330.

Shapiro, Lawrence A. 2013. "Dynamics and Cognition." *Minds & Machines* 23: 353-375.

Shulman, Robert G. 2013. *Brain Imaging: What it Can (and Cannot) Tell us About Consciousness*. Oxford: Oxford University Press.

Sporns, Olaf. 2010. "Brain Networks and Embodiment." In *The Mind in Context*, edited by B. Mesquita, L. F. Barrett, and E. R. Smith, 42–64. New York, NY: Guilford Press.

Thagard, Paul. 2005. *Mind: Introduction to Cognitive Science*. 2nd Edition. Cambridge, MA: The MIT Press.

Thelen, Esther, and Linda B. Smith. 1994. *A Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge, MA: MIT Press.

Thompson, Evan. 2007. *Mind in Life: Biology, Phenomenology, and the Sciences of the Mind*. Cambridge MA: Belknap Press.

Tognoli, Emmanuelle, and J. A. Scott Kelso. 2014. "The Metastable Brain." *Neuron* 81: 35–48.

von Uexküll, Jakob. (1934) 2010. *A Foray into the Worlds of Animals and Humans*. Minneapolis, MN: University of Minnesota Press.

Uttal, William R. 2001. *The New Phrenology: The Limits of Localizing Cognitive Processes in the Brain*. Cambridge, MA: MIT Press.

Uttal, William R. 2011. *Mind and Brain: A Critical Appraisal of Cognitive Neuroscience*. Cambridge, MA: MIT Press.

van Gelder, Tim. 1998. "The Dynamical Hypothesis in Cognitive Science. *Behavioral and Brain Sciences* 21: 615–665.

Van Orden, Guy C., and Heidi Kloos. 2003. "The Module Mistake." *Cortex* 39: 164–166.

Van Orden, Guy C., Bruce F. Pennington, and Gregory O. Stone. 2001. "What do Double Dissociations Prove?" *Cognitive Science* 25: 111–172.

Varela Francisco J., Evan Thompson, and Eleanor Rosch. 1991. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press.

Wilson, Margaret. 2002. "Six Views of Embodied Cognition." *Psychonomic Bulletin & Review* 9: 625–636.

Wilson, Robert A. 1994. "Wide Computationalism." *Mind* 103: 351–372.

Journal of Cognition and Neuroethics

Anesthesia and Consciousness

Rocco J. Gennaro
University of Southern Indiana

Biography

Dr. Rocco J. Gennaro is the Philosophy Department Chairperson and a Professor of Philosophy at the University of Southern Indiana. He received his Ph.D. in 1991 at Syracuse University. Dr. Gennaro's primary research and teaching interests are in Philosophy of Mind/Cognitive Science (especially consciousness), Metaphysics, Early Modern History of Philosophy, NeuroEthics, and Applied Ethics. He has published ten books (as either sole author or editor) and over fifty articles and book chapters in these areas.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). February, 2018. Volume 5, Issue 1.

Citation

Gennaro, Rocco J. 2018. "Anesthesia and Consciousness." *Journal of Cognition and Neuroethics* 5 (1): 49–69.

Anesthesia and Consciousness

Rocco J. Gennaro

Abstract
For patients under anesthesia, it is extremely important to be able to ascertain from a scientific, third-person point of view to what extent consciousness is correlated with specific areas of brain activity. Errors in accurately determining when a patient is having conscious states, such as conscious perceptions or pains, can have catastrophic results. Here, I argue that the effects of (at least some kinds of) anesthesia lend support to the notion that neither basic sensory areas nor the prefrontal cortex (PFC) is sufficient to produce conscious states. I also argue that it this is consistent with and supportive of the higher-order thought (HOT) theory of consciousness. I therefore disagree in some ways with Mehta and Mashour (2013), who argue that evidence from anesthesia mainly favors a first-order representational (FOR) theory, as opposed to HOT theory (and many other theories, for that matter).

Keywords
Anesthesia, Consciousness, Higher-Order Thought, Neural Correlates of Consciousness, Prefrontal Cortex

For patients under anesthesia, it is extremely important to be able to ascertain from a scientific, third-person point of view to what extent consciousness is correlated with specific areas of brain activity (Mashour 2010, Hutt and Hudetz 2015a). Errors in accurately determining when a patient is having conscious states, such as conscious perceptions or pains, can have catastrophic results. Here, I argue that the effects of (at least some kinds of) anesthesia lend support to the notion that neither basic sensory areas nor the prefrontal cortex (PFC) is sufficient to produce conscious states. I also argue that it this is consistent with and supportive of the higher-order thought (HOT) theory of consciousness (Rosenthal 2005, Gennaro 2012). I therefore disagree in some ways with Mehta and Mashour (2013) who argue that evidence from anesthesia mainly favors a first-order representational (FOR) theory as opposed to HOT theory (and many other theories, for that matter).

1. Introduction and Terminology

Perhaps the most fundamental and commonly used notion of "conscious" is captured by Thomas Nagel's famous "what it is like" sense (Nagel 1974). When we are in a conscious mental state, there is "something it is like" for us to be in that state from the subjective or first-person point of view. When we smell a rose or have a conscious

visual experience, there is something it "seems" or "feels like" from our own perspectives. An organism such as a bat is conscious if it is able to experience the world through its echolocation senses. There is also something it is like to be a conscious creature, whereas there is nothing it is like to be a table or tree. This is primarily the sense of "conscious state" used throughout this article. "What it's like" basically means "how a conscious state is for the subject."

Let us also keep in mind the distinction between *state* and *creature* consciousness (Rosenthal 1993). We sometimes speak of an individual mental state, such as a pain or perception, as being conscious. On the other hand, we also often talk about organisms or creatures as conscious, such as when we say that "human beings are conscious" or "cats are conscious." Creature consciousness is simply meant to refer to the fact that an organism is awake, as opposed to sleeping or in a coma. However, some kind of state consciousness is normally implied by creature consciousness; that is, if a creature is conscious, then it must have conscious mental states. Perhaps there are cases where one is state conscious but not creature conscious, such as when one is having a vivid dream. Another possible case is "locked-in syndrome," which is a medical condition where brain damage has affected only motor functions and leaves the patient immobile and unresponsive to stimuli but consciousness remains normal. Mashour and LaRock (2008) refer to locked-in syndrome as the "inverse zombie problem," that is, cases of internally experienced consciousness without any behavioral sign, as opposed to the hypothetical philosopher's "zombie," who is not conscious but behaves in a manner indistinguishable from a conscious human. In this and other related troubling cases, such as persistent vegetative states or minimally conscious states, significant ethical concerns also frequently arise (Braddock 2017).

2. Some Evidence from Anesthesia

The overall available evidence strongly suggests that anesthesia (such as propofol and ketamine) primarily causes the suppression of "feedback" and "top-down" brain mechanisms or connectivity (e.g., Hudetz 2012; Lee et. al. 2013; Crone 2017). It will be helpful to review in some detail the findings and conclusions recently reached by experts in the field. For example, Hudetz explains that "networks based on the posterior parietal-cingulate-precuneus region as a hub and on the nonspecific thalamus are putative candidates for the neural correlate of the state of consciousness [NCCs]" (Hudetz 2012, 299). These areas, he notes, are "prime candidates for the functional networks of the forebrain that play a critical role in maintaining the state of consciousness" (Hudetz 2012,

291). Notice that there is no specific mention of the PFC or basic sensory areas as loci for the NCC.

In a similar fashion, Schrouff et al. (2011, 203) tell us that "results show that deep sedation [due to propofol] was associated with reduced interactions between all… associative cortices…However…the functional interactions of parietal areas were deteriorated to a significantly larger extent than those of frontal or temporal areas." Alkire, Hudetz, and Tononi (2008, 876) explain that "unconsciousness is likely to ensue when a complex of brain regions in the posterior parietal area is inactivated." Thus, there is continued emphasis on parietal brain areas as central to when a subject is having a conscious experience.

Crone et al. (2017) likewise do not focus on the PFC but rather conclude that:

> the data show that loss of consciousness, at least in the context of propofol-induced sedation, is marked by a breakdown of corticopetal projections from the globus pallidus. Effective connectivity between the globus pallidus and the ventral posterior cingulate cortex…fades in the transition from lightly sedated to full loss of consciousness and returns gradually as consciousness recovers. (Crone et al. 2017, 2727)

Further, Hutt and Hudetz (2015b) summarize Blain-Moraes et al. (2014) by explaining that in surgical patients "anesthetic-invariant electroencephalographic effects occur in cortical top-down connectivity. Specifically, ketamine is found to suppress fronto-parietal functional and directional connectivity, similar to that produced by propofol" (Hutt and Hudetz 2015b, 4). They conclude that

> the formerly favored bottom-up mechanisms of anesthetic action focusing on subcortical arousal centers and ascending thalamocortical information transfer are contrasted with the more recent cortical top-down explanations that are inherent to conscious perception and appear to be the preferential target of anesthetic modulation. Substantial electrophysiological and neuroimaging evidence from animal and human investigations supports the top-down mechanisms as a causally sufficient explanation for anesthetic-induced unconsciousness. (Hutt and Hudetz 2015b, 4)

Hudetz and Mashour (2016) frame the matter in the following way:

> After a … dose of propofol, highly connected "hubs" in the brain undergo a reconfiguration, with connectivity patterns in the posterior

> parietal cortex being disrupted. Much like an airport system, a disrupted hub would entail a reduction in incoming traffic, which is exactly what is observed in the form of reduced communication from frontal cortex to posterior parietal cortex. (Hudetz and Mashour 2016, 1233)

Many of the above authors also emphasize the need for necessary connectivity and interaction between different areas of the brain. In addition, Boveroux et al. (2010) show that anesthetic-induced unconsciousness is *not* correlated with inactivation of primary sensory cortical areas. Transverse and sagittal sections of primary visual and the auditory cortices during wakefulness and propofol-induced unconsciousness show the relative preservation across states. That is, neural activity in the primary visual and auditory cortices is preserved while the patient is under anesthesia. Thus, it seems that whatever generally makes a visual or auditory mental state conscious cannot be within the primary cortices. Much the same seems to be the case for various other kinds of conscious states, such as emotions and pains. Indeed, fear and pain, for example, seem to essentially involve an emotional element or cognitive attitude (Baars and Gage 2010, Ch. 13). The limbic system, for example, contains some subcortical structures (such as the amygdala and hypothalamus) as well as some cortical structures (such as the cingulate gyrus). Indeed, the neural realization of pain and emotion is fairly complex and also distributed in different brain areas. In any case, the overall idea is that for normal conscious states, there are "lower" areas of brain activity accompanied higher-level top down cortical interaction. As we have seen, anesthesia mainly disrupts the top-down neural activity.

3. Support for HOT Theory

I think that the above, in turn, provides some support for the view that having conscious states requires having *higher-order thoughts* (HOTs) most often located in-between early sensory areas and the prefrontal cortex (PFC). The HOT theory of consciousness says that what makes a mental state conscious is that there is a suitable higher-order thought directed at the mental state (Rosenthal 2005; Gennaro 2012). HOTs are "meta-psychological" or "metacognitive" states, that is, mental states directed at other mental states. HOT theory is primarily concerned with explaining how conscious mental states differ from unconscious mental states.

Let's back up for a moment. A central question which should be answered by any theory of consciousness is: What makes a mental state a *conscious* mental state? That is, how do we distinguish between unconscious mental states and conscious mental

states? HOT theorists put significant initial weight on what has come to be known as the transitivity principle (TP).

> TP: Conscious states are mental states that I am "aware of" in some sense.

TP seems intuitively true and perhaps even true by definition. When I am in a conscious visual state, I am aware of being in that state. On the other hand, *unconscious* states are those mental states of which I am not aware. I am not aware of being in my current unconscious states. If I am having a subliminal perception, then I am not aware of being in that state. For various reasons, many (including myself) hold that such "meta-awareness" is best understood as a thought composed of concepts, as opposed to, say, a perception.[1]

There is an important and very relevant additional subtlety to HOT theory, however. When a conscious mental state is a first-order world-directed state the higher-order thought (HOT) is *not* itself conscious; otherwise, circularity and an infinite regress would follow. In such cases, we are unaware of the HOTs themselves since our conscious focus in world-directed. But when the HOT is itself conscious, there is a yet higher-order (or third-order) thought directed at the second-order state. In this case, we have *introspection* which involves having a conscious HOT directed at an inner mental state. When one introspects, one's attention is directed back into one's mind, but when one has a first-order conscious state one's attention is outer-focused (see figure 1).

My view is that HOTs, especially the *un*conscious HOTs that accompany first-order conscious states, need not occur in the prefrontal cortex (PFC) area. This also seems to be supported by recent work on anesthesia, as we have already seen to some extent (more on this below). So, although HOT theory demands that conscious states be distributed to some degree in the brain (i.e., beyond basic sensory areas), I opt for a more moderate view with more limited neural connections required (e.g., recurrent feedback loops), especially with respect to first-order conscious states.[2]

1. HOT theory is most often contrasted with HOP (higher-order perception) theory. See e.g. Gennaro (2004), Rosenthal (2004), and Gennaro (2012, chapter 3) for much more discussion of alternative HO theories.

2. Actually, I prefer to treat the lower-order state and the unconscious HOT as parts of a single complex unified state. This is a position I have called the "wide intrinsicality view" or WIV (Gennaro, 1996, 2006, 2012).

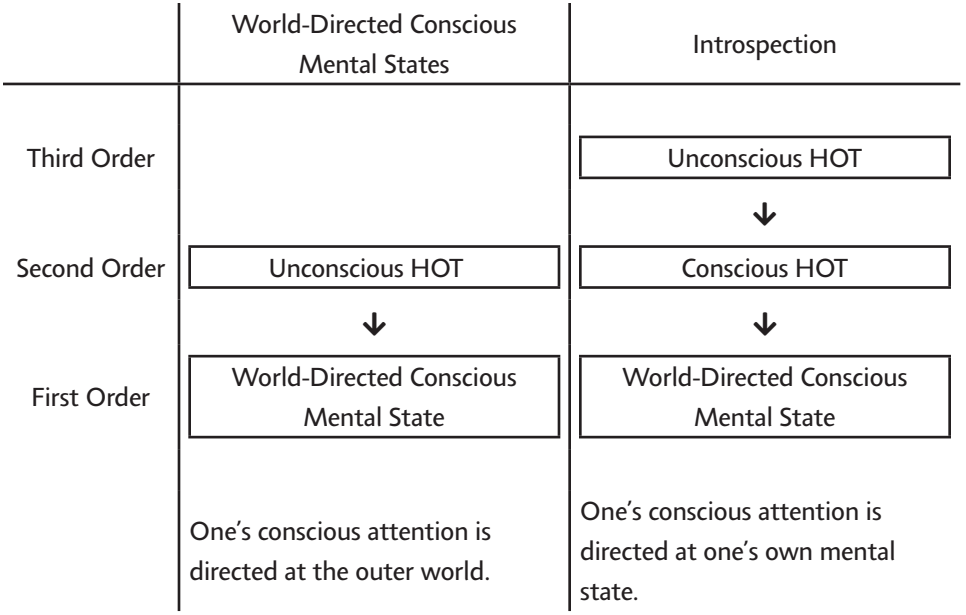


Figure 1. The Higher-Order Thought (HOT) Theory of Consciousness

At the neural level, then, we might borrow from Gerald Edelman and others who have argued that *feedback loops* (or "reentrant pathways" or "back projections") in the neural circuitry of the brain are essential for conscious awareness (Edelman and Tononi 2000a; 2000b). As Patricia Churchland once put it, "The idea is that some neurons carry signals from more peripheral to more central regions…while others convey more highly processed signals in the reverse direction. . .It is a general rule of cortical organization that forward-projecting neurons are matched by an equal or greater number of back-projecting neurons" (2002, 148–149). The brain structures involved in loops seem to resemble the structure of at least some form of HOT theory; namely, lower-order and higher-order states mutually interacting to produce conscious states. Edelman and Tononi, for example, sometimes emphasize the global nature of conscious states, and it is reasonable to interpret this as the view that conscious states comprise both the higher-and lower-order states. What they call the "dynamic core" is generally "spatially distributed and thus cannot be localized to a single place in the brain" (Edelman and Tononi 2000a, 146).

Other related support comes from Victor Lamme (2003; 2004) who argues that recurrent processing is necessary before the properties of an object are attentively grouped and the stimulus can enter consciousness. Based on experimental results, such as texture segregation and visual search tasks, Lamme argues that the so-called "feedforward sweep" is not sufficient for consciousness. Lamme also explains that "backward masking" renders a visual stimulus invisible by presenting a second stimulus shortly after the first (about 40 milliseconds later, but perhaps up to 110 msecs). Nonetheless, the masked (invisible) stimulus still evokes significant feedforward activation in visual and even nonvisual areas. It seems that the feedback interaction from higher to lower visual areas is suppressed by backward masking, thereby disrupting reentrant processing and inhibiting the production of a conscious states (Fahrenfort, Scholte, and Lamme 2007; Kouider and Dehaene 2007). This suggests that neural activity beyond basic sensory areas is necessary in order to have a conscious perceptual state.

To use one nonvisual example, consider tactile awareness in the somatosensory cortex, extensively reviewed in Gallace and Spence (2010). Once again, there seems to be evidence of feedback activity from higher brain areas necessary for conscious tactile experiences. Gallace and Spence explain that "activation of early sensory areas is insufficient to sustain awareness of tactile sensations. . . . Higher order structures seem necessary" (2010, 50). So, we might say, that tactile information becomes conscious when earlier somatosensory areas trigger a feedback signal from a higher-order representation.

4. Mehta and Mashour's Argument

I therefore disagree to some extent with Mehta and Mashour (2013; M&M hereafter), who argue that evidence from anesthesia mainly favors a first-order representational (FOR) theory, as opposed to HOT theory (and many other theories, for that matter). A FOR theory of consciousness is one that attempts to explain and reduce conscious experience primarily in terms of world-directed (or first-order) intentional states (e.g. Tye 2000, Dretske 1995).

M&M start out by noting that

> a complete theory of consciousness must explain…*first* …what we term *general consciousness*: what makes a state conscious *at all*, as opposed to wholly unconscious. This should explain the difference between, e.g., one's conscious state when one sees a red thing and one's unconscious state when one is anesthetized…. *Second*…*specific consciousness*: what gives a state its specific *phenomenal character* [or

> "content"], rather than some alternative phenomenal character. This… should explain the difference between one's conscious state when one sees something red and one's conscious state when one sees something green (or hears a loud noise, or feels pain). (Mehta and Mashour 2013, 2)

M&M also define two broad types of systems that participate in conscious processing, namely

> *sensory systems* [which] are dedicated to the detection of highly specific perceptible features [which] may be tuned to modality-specific properties (such as color and tone) or properties detectible via multiple modalities (such as motion and spatial location)…[and] *post-sensory systems* [which] perform a broad variety of functions, including modulation of sensory processing via top-down attention... (Mehta and Mashour 2013, 2)

But then, M&M tie HOTs mainly to the prefrontal cortex (PFC) when they say the following:

> Lau and Rosenthal (2011) hypothesize that higher-order representations are harbored in post-sensory systems, such as posterior parietal regions and *especially the dorsolateral prefrontal cortex*, while first-order representations are harbored in sensory systems. Although this neural interpretation is not forced on higher-order representationalists, we will henceforth adopt it because (a) several higher-order representationalists have endorsed this interpretation, and (b) higher-order representationalism is very difficult to test scientifically without some neural interpretation. Given this interpretation, higher-order representationalists identify post-sensory areas as the neural correlates of *general* consciousness; they identify post-sensory and perhaps also sensory areas as the neural correlates of *specific* consciousness (Mehta and Mashour 2013, 3, my emphases).

So, it turns out that one main reason that M&M favor FOR over HOT theory is that they suppose that all (most?) HOT theorists think that HOTs mainly occur in "post sensory systems…*especially the dorsolateral prefrontal cortex*." But, when we look closely at M&M's own paper, we can see that much of the evidence they cite, at minimum, equally favors HOT theory and the same surely goes for the other anesthesia evidence

cited earlier. That is, anesthesia primarily targets post-sensory areas but not the PFC. One would expect HOTs to occur in post-sensory areas. To be fair, M&M do ultimately concede that a modified HOT theory could accommodate the data.

As M&M rightly acknowledge, and as we saw earlier, there is other supportive evidence from Lamme and colleagues. M&M ask us to

> consider visual processing, which includes both a fast *feedforward sweep* and slower *recurrent processing* loops. In the feedforward sweep, which is completed in about 100ms, activation proceeds in a swift, unidirectional cascade from the retina, to the lateral geniculate nucleus of the thalamus, to V1 in the occipital cortex, to higher visual processing areas (includingV2, V3, V4, and V5), and finally to more rostral structures. By contrast, recurrent processing loops involve reciprocal information transfer between the cortical areas, and corresponding thalamic regions via horizontal and feedback connections. Recurrent processing may occur at many scales, from a local scale (within a sensory modality) to a global scale (implicating executive function or spanning different sensory modalities). (Mehta and Mashour 2013, 2-3)

Still, M&M do rightly suppose that "…the simplest explanation of the current data is that…post-sensory processing regions *alone* are the neural correlates of general consciousness. [But then they say that] "this result favors first-order representationalism and higher-order representationalism" (Mehta and Mashour 2013, 6). It seems to me, however, that it favors the higher-order approach much more so than FOR. As a matter of fact, Mashour (2014) elsewhere himself says that "there is growing evidence… that general anesthetics disrupt higher-order cognitive processes and that networks of association cortex may be particularly susceptible to anesthetic effects…" (Mashour 2014, 7). Similarly, he explains that "consciousness and anesthetic-induced unconsciousness are associated with multimodal association cortex rather than primary sensory cortex" (Mashour 2014, 2).

It is somewhat unclear to me how exactly a FOR theorist can suppose that post-sensory areas constitute the locus of "general" consciousness. Presumably, it is because M&M understand FOR theories to include the notion that first-order conscious states must be "available to" post-sensory areas, that is, a conscious representation must be "available to" the subject as a reason for action and belief formation. They explain as follows:

> according to [FOR] theory, consciousness is hypothesized to consist in (i) first-order representations directed at the world which (ii) are directly available to the subject for action selection, belief formation, planning, etc. Condition (i) embodies an approach to specific consciousness: the specific phenomenal character of a representation is determined wholly by its content. Condition (ii) complements this with an approach to general consciousness: for a representation to be conscious rather than unconscious is for it to be directly available to the subject for action selection, belief formation, planning, etc. (Mehta and Mashour 2013, 6).

But it is difficult to see how the mere "availability" to the subject for various purposes can make a representation conscious, at least as opposed to an *actual* HOT directed at the first-order representation. Similarly, it is unclear just how such a disposition can confer *actual* consciousness on an otherwise unconscious mental state. Further, and more to the point, when an anesthetic suppresses post-sensory activity, it is so much clearer that HOTs cannot occur in those areas and thus conscious states will not occur. On the other hand, what does it mean to say that the "availability" to the subject has been suppressed? How does an unconscious first-order representation "know" that such availability has been cut off? Isn't the first-order state still "potentially" conscious in some sense? Why aren't sensory areas enough to produce conscious states if sensory area mental states are merely disposed to be made aware of by post-sensory states?[3]

It is worth briefly noting here that Flohr (2000) has argued that N-methyl-D-aspartate (NMDA)-mediated transient neural assemblies are essential to consciousness based on evidence from anesthesia. Flohr also cites HOT theory approvingly as a way to explain the overall structure of conscious states and the effects of anesthesia. The idea is that anesthetics destroy conscious mental activity because they interfere with the functioning of NMDA receptors. According to Flohr, the activation of the NMDA system is necessary for the mechanisms underlying consciousness. Flohr explicitly relates his theory to higher-order accounts of consciousness by arguing that the NMDA synapse implements the binding mechanism that the brain uses to produce widely distributed representations (Flohr 2000, 252–253). One potential problem with Flohr's account might be that he is focused too narrowly on overall creature consciousness in the sense of

3. Thanks to Benedicte Veillet for pressing this point. For more on these themes and different FOR and HOT theories, see Rosenthal 2004 and Gennaro 2012, chapter 3.

a creature being awake or aware of its surroundings (as opposed to state consciousness). In some ways, however, his emphasis makes sense when thought of from the point of view of anesthesia and neurochemistry. The question is indeed often about whether or not the *patient* is unconscious or when the *person* loses consciousness. Of course, we still want to know if the patient is experiencing any conscious states, and especially pains.

5. HOT Theory and the PFC

If we are correct thus far and anesthesia does not mainly target PFC areas, it would seem that a HOT theorist should look elsewhere for where at least many HOTs occur in the brain. Thus, I have argued elsewhere (Gennaro 2012, ch. 9) that HOT theory need not be committed to the view that the PFC is required for having conscious states, contrary to Kriegel (2007) and Block (2007) who (like Lau and Rosenthal) also suppose that HOT theorists hold that HOTs are realized in the PFC. Still, it is likely true that the PFC is required for the more sophisticated *introspective* states, but this is not a problem for HOT theory because it does not require introspection for merely having first-order (outer-directed) conscious states (see figure 1 again).

On Kriegel's theory, for example, three "elements" are required for NCCs:

1. A "floor-level" (or first-order) representation,

2. A "higher-order" representation of (1), and

3. The "functional integration" of (1) and (2) into a single unified state via some binding mechanism.

The likely NCCs for the floor-level representations will depend on the modality, such as V1-V5/MT for perceiving a moving patch of blue color. According to his view, this has to do with the *contents* of consciousness (or "specific" consciousness, using M&M's terminology), as opposed to consciousness *as such* (or "general" consciousness). So far so good, but then Kriegel says that the likely NCCs for the second-level or higher-order representations are in the PFC. In reply, we should point out that Kriegel's discussion of his "second element" reflects some very sophisticated abilities, such as "executive functions" and "attentional control," which are better understood as *introspective* capacities. Again, it might very well be that *these* higher cognitive capacities are indeed subserved by PFC activity but there is no reason to think that they are required for having first-order conscious mental states (even according to HOT theory).

Block (2007, 485) also states that "since frontal areas are likely to govern higher-order thought, low frontal activity in newborns [and most animals?] may well indicate a lack of higher order thoughts about genuine sensory experiences." Although I agree with Block that PFC activity is not necessary for having first-order conscious states, I disagree with the claim that "frontal areas are likely to govern higher-order thought," unless he primarily means *introspection*; that is, conscious HOTs. In short, a HOT theorist is not be committed to the view that PFC activity is required for having all conscious states. This is also important in order to counter the frequently made charge that HOT theory rules out infant and (most) animal consciousness (Seager 2004). Indeed, one HOT theorist has accepted this otherwise undesirable consequence of HOT theory (Carruthers 2000, 2005).[4] It is also clear from the evidence adduced earlier that neural activity in post-sensory areas (but not including the PFC) result in first-order conscious states and are responsible for general consciousness. After all, it is suppression of activity in those areas which eliminate consciousness.

In addition to the anesthesia-based evidence cited thus far, there are independent reasons to think that conscious states occur without the PFC. For example, conscious experience is not eliminated entirely when there is extensive PFC damage, even in lobotomies (Pollen 2003). And when subjects are engaged in a perceptual task or absorbed in watching a movie, there is widespread neural activation but little PFC activity (Goldberg, Harel, and Malach 2006). Although other studies do show *some* PFC activation in similar experiments, this is mainly because of the need for subjects to *report* their experiences and the PFC is likely to be activated when there is *reflection* or *introspection* about one's experiences.

But is there any positive reason to think that unconscious HOTs can occur outside of the PFC? I think there is. Assuming that HOTs can be understood as a form of self-consciousness, as seems reasonable I think, unconscious HOTs might then be regarded as a kind of "pre-reflective" self-consciousness (as opposed to reflective or introspective self-consciousness). Newen and Vogeley (2003), for example, go so far as to distinguish five levels of self-consciousness ranging from "phenomenal self-acquaintance" and "conceptual self-consciousness" up to "iterative meta-representational self-consciousness." Citing numerous experiments, they point to various "neural signatures" of self-consciousness, but the PFC is rarely mentioned, and then, usually only with regard to the more sophisticated forms of self-consciousness. Other brain areas are much more prominently

4. But see e.g. Gennaro 2009, Gennaro 2012, chapters 7 and 8, for my most recent attempts to rebut that line of argument.

identified, such as the medial and inferior parietal cortices, the temporo-parietal cortex, the posterior cingulate cortex, and the anterior cingulate cortex.

Even when considering the neural signatures of "theory of mind" and "mindreading," Newen and Vogeley cite experiments indicating that such meta-representation is best located in the anterior cingulate cortex and also showed activation in the right temporo-parietal junction and the medial aspects of the superior parietal lobe. Related, and more recent, support for this position can be found, for example, in the work of Rebecca Saxe, who has extensively studied brain regions most associated with thinking about other people's thoughts, sometimes called "mindreading" and involving a so-called "theory of mind" (Saxe 2009; 2010). The temporo-parietal junction, the posterior cingulate, the medial precuneus, and parts of the temporal sulcus are identified as the primary sites for this kind of cognition.

6. Worries about NCCs?

One might wonder if we are playing too fast and loose with talk of NCCs. At the least, it is important to avoid several problems and potential pitfalls when discussing evidence related to NCCs. For example, one issue is determining exactly how the NCC is related to consciousness. Although a case can be made that many NCCs are *necessary* for conscious mentality, it is sometimes unclear if they are *sufficient*. For one thing, many candidates for NCCs can also occur unconsciously, such as feedback loops in earlier sensory areas. Second, there are obviously other background conditions that must obtain (e.g., breathing, proper blood flow) in order for a given NCC to suffice for consciousness. Even pinning down a narrow-enough necessary condition is not as easy as it might seem.

A related worry has to do with the very use of the term "correlate." As any philosopher, scientist, and even undergraduate student should know, saying that "A is correlated with B" is rather weak by itself (though it can be an important first step), especially if one wishes to establish a stronger *causal* or *identity* claim between consciousness and neural activity. Even if a solid correlation can be established, we cannot automatically conclude that there is an identity relation. One might even suppose that the search for NCCs is somewhat neutral with respect to the metaphysics of mind, though a materialist might urge us at some point to accept an identity claim on the basis of the principle of simplicity. Still, perhaps A causes B or B causes A, and that's why we find the correlation. Maybe there is even some *other* neural process C that causes both A and B.

Chalmers (2000) presents several useful distinctions and definitions for the purpose of conceptual clarity (cf. Block 2007; Hohwy 2007). For one thing, we should distinguish

between having the conscious mental state itself (or "vehicle") and its content. Thus, Chalmers presents the following definitions:

> A *content* NCC is a neural representational system N such that the content of N directly correlates with the content of consciousness. (Chalmers 2000, 20; italics mine).
>
> A *state* N1 of system B is a neural correlate of phenomenal property P if N's being in N1 directly correlates with the *subject* having P. (22; italics mine).

In our discussion of anesthesia, we must be careful not to suppose that post-sensory brain areas determine what M&M call "specific" consciousness; that is, the *content* NCCs of conscious perceptual states. Actually, some HOT theorists think that it takes *both* sensory and post-sensory areas to produce *specific* consciousness in the sense that the first-order content must be properly referenced or matched by the HOT's content (Gennaro 2012).[5] In any case, we have already seen that general consciousness is not realized in basic sensory areas and is likely to be in various post-sensory areas. This is more like what Chalmers has in mind by "state" NCC as opposed to the "content" NCC.

It is then important to recognize that any interesting NCC would at least need to isolate the *minimal* brain area responsible for a conscious state. Thus, one finds the following:

> An NCC is a *minimal neural system* N such that there is a mapping from states of N to states of consciousness, where a given state of N is sufficient, under conditions C, for the corresponding state of consciousness. (Chalmers 2000, 31; italics mine).

From the evidence we have examined regarding anesthesia, it certainly seems as if the minimal neural system N in question would include both sensory and post-sensory areas when one is having a conscious first-order perceptual state. Each area is necessary for conscious states, but they are jointly sufficient for the relevant state of consciousness. To be more specific with regard to post-sensory areas, the posterior parietal and cingulate cortices seem to be especially important.

5. There is significant disagreement on this matter among HOT theorists, such as what would happened if or when a HOT's content misrepresents its target mental state or even when there is no target state at all. See Gennaro (2012, chapters 4 and 6) for some discussion.

Others make similar remarks and distinctions with respect to NCCs, such as when Block (2007, 489) explains that a "minimal neural basis is a necessary part of a neural sufficient condition for conscious experience," and when Koch (2004, 16) tells us that the NCC is "the minimal set of neuronal events and mechanisms jointly sufficient for a specific conscious percept." The main point is to find a neural correlation that is a reasonably interesting subset of the entire brain activity at a given time. It would be much less informative, and perhaps just trivial, to learn that the *entire* brain is sufficient for having a conscious state. In a similar vein, one might distinguish between the *core* and *total* NCC. The *core* neural basis of a conscious state is the part of the total neural basis that distinguishes conscious states from states with other conscious contents. This is very much like what M&M call "specific" consciousness and what Chalmers calls the "content NCC." The *total* neural basis of a conscious state is itself sufficient for the instantiation of that conscious state (Block 2007, 482). Once again, this seems to involve both sensory and post-sensory areas. We also need to distinguish the NCC from what might be viewed as other "enabling conditions," which refer to other aspects of a functioning body, such as proper blood flow and functioning lungs and heart (see Block 2007, 485-486, for some discussion). There may be deeper problems here as well and perhaps we even need new experimental approaches (Hohwy 2009). At the least, it is also crucial to design experiments with controls such that the *only* difference between a pair of trials is the presence of consciousness. We can then use, say, functional magnetic resonance imaging (fMRI) to ascertain any neural differences between such cases. Neurophysiologists sometimes use the "subtraction method;" namely, subtract control or constant background neural activity from the neural activity of a given task which involves conscious perception.

7. Conclusion

The effects of (at least some kinds of) anesthesia lend support to the notion that neither basic sensory areas nor the prefrontal cortex (PFC) is sufficient to produce conscious states. Rather, it takes the combination of sensory areas and post-sensory areas (not including the PFC) in order for there to be a first-order conscious state. It was also argued that it this is consistent with and most supportive of the higher-order thought (HOT) theory of consciousness. Still, we must be careful as to how to characterize the NCCs in question.

References

Alkire, Michael, Anthony Hudetz, and Giulio Tononi. 2008. "Consciousness and Anesthesia." *Science* 322 (5903): 876–880.

Baars, Bernard, and Nicole Gage. 2010. *Cognition, Brain, and Consciousness, 2nd edition*. San Diego: Elsevier/Academic Press.

Blain-Moraes, Stefanie, UnCheol Lee, SeungWoo Ku, GyuJeong Noh, and George Mashour. 2014. "Electroencephalographic Effects of Ketamine on Power, Cross-frequency Coupling, and Connectivity in the Alpha Bandwidth." *Frontiers in Systems Neuroscience* 8 (July): 1-9.

Block, Ned. 2007. "Consciousness, Accessibility, and the Mesh between Psychology and Neuroscience." *Behavioral and Brain Sciences* 30 (5): 481-499.

Boveroux, Pierre, Audrey Vanhaudenhuyse, Marie-Aurelie Bruno, Quentin Noirhomme, S. Lauwick, and A. Luxen, Christian Degueldre, Alain Plenevaux, Caroline Schnakers, Christophe Phillips, Jean-Francois Brichant, Vincent Bonhomme, Pierre Maquet, Michael D. Greicius, Steven Laureys, and Melanie Boly. 2010. "Breakdown of Within- and Between-Network Resting State Functional Magnetic Resonance Imaging Connectivity during Propofol-Induced Loss of Consciousness." *Anesthesiology* 113 (5): 1038-1053.

Braddock, Matthew. 2017. "Should We Treat Vegetative and Minimally Conscious Patients as Persons?" *Neuroethics* 10 (2): 267-280.

Carruthers, Peter. 2000. *Phenomenal Consciousness*. Cambridge: Cambridge University Press.

Carruthers, Peter. 2005. *Consciousness: Essays from a Higher-Order Perspective*. New York: Oxford University Press.

Chalmers, David. 2000. "What is a Neural Correlate of Consciousness?" In *Neural Correlates of Consciousness: Empirical and Conceptual Questions*, edited by Thomas Metzinger, 17-39. Cambridge: The MIT Press.

Churchland, Patricia. 2002. *Brain-Wise*. Cambridge: The MIT Press.

Crone, Julia, Evan Lutkenhoff, Branden Bio Steven Laureys, and Martin Monti. 2017. "Testing Proposed Neuronal Models of Effective Connectivity within the Cortico-basal Ganglia-thalamo-cortical Loop During Loss of Consciousness." *Cerebral Cortex* 27 (4): 2727-2738.

Dretske, Fred. 1995. *Naturalizing the Mind*. Cambridge: The MIT Press.

Edelman, Gerald, and Giulio Tononi. 2000a. "Reentry and the Dynamic Core: Neural Correlates of Conscious Experience." In *Neural Correlates of Consciousness: Empirical and Conceptual Questions*, edited by Thomas Metzinger, 139-151. Cambridge: The MIT Press.

Edelman, Gerald, and Giulio Tononi. 2000b. *A Universe of Consciousness*. New York: Basic Books.

Fahrenfort, Johannes, Steven Scholte, and Victor Lamme. 2007. "Masking Disrupts Reentrant Processing in Human Visual Cortex." *Journal of Cognitive Neuroscience* 19 (9): 1488-1497.

Flohr, Hans. 2000. "NMDA Receptor-Mediated Computational Processes and Phenomenal Consciousness." In *Neural Correlates of Consciousness: Empirical and Conceptual Questions*, edited by Thomas Metzinger, 245-258. Cambridge: The MIT Press.

Gallace, Alberto, and Charles Spence. 2010. "Touch and the Body: The Role of the Somatosensory Cortex in Tactile Awareness." *Psyche* 16 (1): 31-67.

Gennaro, Rocco. 1996. *Consciousness and Self-Consciousness: A Defense of the Higher-Order Thought Theory of Consciousness*. Amsterdam: John Benjamins.

Gennaro, Rocco. 2004. *Higher-Order Theories of Consciousness: An Anthology*. Amsterdam: John Benjamins.

Gennaro, Rocco. 2006. "Between Pure Self-Referentialism and the (Extrinsic) HOT Theory of Consciousness." In *Self-Representational Approaches to Consciousness*, edited by Uriah Kriegel and Ken Williford, 221-248. Cambridge: The MIT Press.

Gennaro, Rocco. 2009. "Animals, Consciousness, and I-thoughts." In *Philosophy of Animal Minds*, edited by Robert Lurz, 184-200. New York: Cambridge University Press.

Gennaro, Rocco. 2012. *The Consciousness Paradox: Consciousness, Concepts, and Higher-Order Thoughts*. Cambridge: The MIT Press.

Goldberg, Ilan, Michael Harel, and Rafael Malach. 2006. "When the Brain Loses Its Self: Prefrontal Inactivation during Sensorimotor Processing." *Neuron* 50 (2): 329-339.

Guldenmund, Pieter, Ithabi S. Gantner, Katherine Baquero, Tushar Das, Athena Demertzi, Pierre Boveroux, Vincent Bonhomme, Audrey Vanhaudenhuyse, Marie-Aurélie Bruno, Olivia Gosseries, Quentin Noirhomme, Muriëlle Kirsch, Mélanie Boly, Adrian M. Owen, Steven Laureys, Francisco Gómez, and Andrea Soddu. 2016. "Propofol-Induced Frontal Cortex Disconnection: A Study of Resting-State Networks,

Total Brain Connectivity, and Mean BOLD Signal Oscillation Frequencies." *Brain Connectivity* 6 (3): 225-237.

Hohwy, Jakob. 2007. "The Search for Neural Correlates of Consciousness." *Philosophy Compass* 2 (3): 461-474.

Hohwy, Jakob. 2009. "The Neural Correlates of Consciousness: New Experimental Approaches Needed?" *Consciousness and Cognition* 18 (2): 428-438.

Hudetz, Anthony. 2012. "General Anesthesia and Human Brian Connectivity." *Brain Connectivity* 2 (6): 291-302.

Hudetz, Anthony and George Mashour. 2016. "Disconnecting Consciousness: Is There a Common Anesthetic End Point?" *Anesthesia and Anelgesia* 123 (5): 1228-1240.

Hutt, Axel and Anthony Hudetz. eds. 2015a. *General Anesthesia: From Theory to Experiments*. Lausanne: Frontiers Media (*Frontiers in Systems Neuroscience*).

Hutt, Axel and Anthony Hudetz. 2015b. "Editorial: General Anesthesia: From Theory to Experiments." In *General Anesthesia: From Theory to Experiments*, edited by Axel Hutt and Anthony Hudetz, 4-6. Lausanne: Frontiers Media (*Frontiers in Systems Neuroscience*).

Koch, Christof. 2004. *The Quest for Consciousness: A Neurobiological Approach*. Englewood, CO: Roberts.

Kouider, Sid, and Stanislas Dehaene. 2007. "Levels of Processing During Non-conscious Perception: A Critical Review of Visual Masking." *Philosophical Transactions of the Royal Society of London, B* 362 (1481): 857-875.

Kriegel, Uriah. 2007. "A Cross-Order Integration Hypothesis for the Neural Correlate of Consciousness." *Consciousness and Cognition* 16 (4): 897-912.

Lamme, Victor. 2003. "Why Visual Attention and Awareness Are Different." *Trends in Cognitive Sciences* 7 (1): 12-18.

Lamme, Victor. 2004. "Separate Neural Definitions of Visual Consciousness and Visual Attention: A Case for Phenomenal Awareness." *Neural Networks* 17 (5): 861-872.

Lau, Hakwan, and David Rosenthal. 2011. "Empirical Support for Higher-Order Theories of Conscious Awareness." *Trends in Cognitive Sciences* 15 (8): 365-373.

Lee, UnCheol, SeungWoo Ku, GyuJeong Noh, SeungHye Baek, ByungMoon Choi, and George Mashour. 2013. "Disruption of Frontal-Parietal Communication by Ketamine, Propofol, and Sevoflurane." *Perioperative Medicine* 118 (6): 1264-1276.

Mashour, George. ed. 2010. *Consciousness, Awareness, and Anesthesia*. Cambridge: Cambridge University Press.

Mashour, George. 2014. "Top-down Mechanisms of Anesthetic-Induced Unconsciousness." *Frontiers of Systems Neuroscience* 8 (June): 1-10.

Mashour, George, and Eric LaRock. 2008. "Inverse Zombies, Anesthesia Awareness, and the Hard Problem of Unconsciousness." *Consciousness and Cognition* 17 (4): 1163-1168.

Mehta, Neil and George Mashour. 2013. "General and Specific Consciousness: A First-Order Representationalist Approach." *Frontiers in Psychology* 4 (July): 1-9.

Nagel, Thomas. 1974. "What is it Like to be a Bat?" *Philosophical Review* 83 (4): 435-456.

Newen, Albert, and Kai Vogeley. 2003. "Self-Representation: Searching for a Neural Signature of Self-Consciousness." *Consciousness and Cognition* 12 (4): 529-543.

Pollen, Daniel. 2003. "Explicit Neural Representations, Recursive Neural Networks and Conscious Visual Perception." *Cerebral Cortex* 13 (8): 807-814.

Rosenthal, David. 1993. "State Consciousness and Transitive Consciousness." *Consciousness and Cognition* 2 (3): 355-363.

Rosenthal, David. 2004. "Varieties of Higher-Order Theory." In *Higher-Order Theories of Consciousness: An Anthology*, edited by Rocco Gennaro, 17-44. Amsterdam: John Benjamins.

Rosenthal, David. 2005. *Consciousness and Mind*. New York: Oxford University Press.

Saxe, Rebecca. 2009. "Theory of Mind (Neural Basis)." In *Encyclopedia of Consciousness*, edited by William Banks. Boston: Elsevier/Academic Press.

Saxe, Rebecca. 2010. "The Right Temporo-Parietal Junction: A Specific Brain Region for Thinking about Thoughts." In *Handbook of Theory of Mind*, edited by Alan Leslie and Tamsin German. New York: Routledge.

Schrouff, Jessica, Vincent Perlbarg, Mélanie Boly, Guillaume Marrelec, Pierre Boveroux, Audrey Vanhaudenhuyse, Marie-Aurélie Bruno, Steven Laureys, Christophe Phillips, Mélanie Pélégrini-Issac, Pierre Maquet, Habib Benali. 2011. "Brain Functional Integration Decreases During Propofol-Induced Loss of Consciousness." *NeuroImage* 57 (1): 198-205.

Seager, William. 2004. "A Cold Look at HOT Theory." In *Higher-Order Theories of Consciousness: An Anthology*, edited by Rocco Gennaro, 255-275. Amsterdam: John Benjamins.

Tye, Michael. 2000. *Consciousness, Color, and Content*. Cambridge: The MIT Press.

Journal of Cognition and Neuroethics

Decision-Theoretic Consequentialism and the Desire-Luck Problem

Sahar Heydari Fard
University of Cincinnati

Biography
Sahar Heydari Fard is a PhD candidate in philosophy at University of Cincinnati. The focus of her work is on moral epistemology and the morality of social movements.

Publication Details
Journal of Cognition and Neuroethics (ISSN: 2166-5087). February, 2018. Volume 5, Issue 1.

Citation
Heydari Fard, Sahar. 2018. "Decision-Theoretic Consequentialism and the Desire-Luck Problem." *Journal of Cognition and Neuroethics* 5 (1): 71–84.

Decision-Theoretic Consequentialism and the Desire-Luck Problem

Sahar Heydari Fard

Abstract

Jackson (1991) proposes an interpretation of consequentialism, namely, the Decision Theoretic Consequentialism (DTC), which provides a middle ground between internal and external criteria of rightness inspired by decision theory. According to DTC, a right decision either leads to the best outcomes (external element) or springs from right motivations (internal element). He raises an objection to fully external interpretations, like objective consequentialism (OC), which he claims that DTC can resolve. He argues that those interpretations are either too objective, which prevents them from giving guidance for action, or their guidance leads to wrong and blameworthy actions or decisions. I discuss how the emphasis on blameworthiness in DTC constraints its domain to merely the justification of decisions that relies on rationality to provide a justification criterion for moral decisions. I provide examples that support the possibility of rational but immoral decisions that are at odds with DTC's prescription for right decisions. Moreover, I argue what I call the desire-luck problem for the external element of justification criterion leads to the same objection for DTC that Jackson raised for OC. Therefore, DTC, although successful in response to some objections, fails to provide a prescription for the right decision.

Keywords

Decision Theory, Consequentialism, Moral Luck, Desires, Emotion, Rationality

Introduction

In his 1991 paper, "Decision-Theoretic Consequentialism and the Nearest and Dearest Objection," Frank Jackson responds to Bernard William's objection that there is a tension between consequentialism and some of our fundamental intuitions. Jackson's main project is to show that a proper understanding of consequentialism, namely decision-theoretic consequentialism (DTC), resolves that seeming tension. He argues against the already existing interpretations of consequentialism by proposing a dilemma: they are either too objective, which prevents them from giving guidance for action, or their guidance leads to wrong actions. An example of such an interpretation is Peter Railton's 1984 proposal of objective consequentialism (OC), in which the only criterion of rightness is fully external, whether *in fact* the action maximized utility or not.

Inspired by Thomas Nagel's 1979 problem of moral luck, which I explain shortly, Jackson moves away from the external criterion of rightness toward an internal criterion that is compatible with consequentialism. Jackson argues that "the fact that an action

might have the best results might be obscure for the agent…. Hence, the fact that a course of action would have the best results is not in itself a guide to action, for a guide to action must in some appropriate sense be present to the agent's mind" (Jackson 1991, 466). Jackson concludes that a theory of right motive should supplement the criterion of rightness. He states, "I thus am agreeing with Thomas Nagel's claim that morality requires of us not only certain forms of conduct but also the motives required to produce the conduct" (Jackson 1991, 468). In Jackson's view, either a right motive or maximization of utility can make an action right, but neither are necessary. He provides an account of right motivation as a combination of beliefs and right desires. Beliefs as the internal element of right motivation are the subjective probability function that determines the agent's degree of credence for the occurrence of different outcomes. And desires as the external element of right motivation ought to conform to the consequentialist ranking of the alternatives.

Jackson's proposal, although successful in responding to some objections, leads to similar problems that motivated it. I suggest a distinction between the heuristic and justificatory role of decisions in the discussion of the criterion of rightness. According to this distinction, DTC and OC are on the same page when it comes to the heuristic role of decisions and of the decision-making process. They both consider the determination of the right decision process highly contextual and a matter of empirical evidence. Thus, the advantage of DTC is that it provides a prescription for a justified decision, a decision that, regardless of the outcomes, is not blameworthy. A justified decision, according to DTC, is a rational decision, which is defined by decision theory. However, I provide examples in which a rational decision is morally unjustified and therefore blameworthy. The possibility of such cases crumbles the hope of providing a criterion for moral justification based on a decision-theoretic criterion of rationality.

DTC distinguishes itself from fully internal views about right motivation by including an external criterion of rightness for desires. However, this inclusion makes a new problem that I call the desire-luck problem. According to this problem, it seems justified to morally assess people's desires when moral assessment is appropriate for matters over which the agent has control, but I discuss studies that show the significant influence of outside factors on our desires in the process of decision making. Therefore, the rightness of a motivation and thereby justification of the related decision can be dependent upon factors over which the agent does not have any control. Thus, DTC is facing a familiar dilemma. If it does not provide any prescription for how to achieve the right desire, its criterion for the rightness of motivation is questioned by the desire-luck problem. But

if DTC provides a prescription for achieving the right desire, then there might be cases in which following those prescriptions leads to unjustified decisions in another context.

Decision-Theoretic Consequentialism

Nagel's moral luck problem inspires Jackson's project. The problem of moral luck concerns the contradiction between two moral intuitions. The intuition that "moral assessment is only appropriate for matters over which the agent has control" does not seem compatible with the intuition that "it is sometimes justified to assign praise or blame to things over which agents do not have full control" (Jackson 1991). Jackson reformulates this problem and introduces a new one that I call the prescription problem. If I understand Jackson correctly, the idea is that a moral theory cannot prescribe something that it considers blameworthy. In other words, assuming that wrongness and blameworthiness have to be co-extensional,[1] it is contradictory to prescribe something that might turn out to be wrong. Agents do not always have all the information about what turns out to be right. Thus, a moral theory cannot blame an agent for making a decision that leads to a wrong outcome when the agent did not know what exactly the outcome would be. This is how Jackson explains the problem:

> When we act, we must perforce use what is available to us at the time, not what may be available to us in the future or what is available to someone else, and least of all not what is available to a God-like being who knows everything about what would, will, and did happen. (Jackson 1991, 472)

The prescription problem emphasizes the lack of direct relationship between the morally right action and the morally right decision. We cannot be certain about the outcome of our decision, but this does not imply that we are not responsible for making the right decision. Thus, there should be a criterion of right decision that guarantees that, regardless of the outcome, making a decision based on that criterion does not make the agent blameworthy. Considering this problem, the fact that OC defines rightness only in terms of the external outcome leads to a dilemma. If OC does not provide any prescription for a right decision, then the blame or praise worthiness of a decision seems arbitrary and a matter of moral luck. But, if OC prescribes the right decision,

1. This is not an obvious assumption and not everyone agree with it. But it seems necessary to make this assumption to understand Jackson's argument.

then there might be cases in which those prescriptions lead to worse outcomes that are blameworthy:

> Suppose that consequentialism says nothing about the mind of the agent at all. It says merely that right action is action with property φ, for some consequentialist treatment of φ which pertains solely to what in fact would happen and not at all to what the agent thinks. In that case, consequentialism, as Williams puts it, "has to vanish from making any distinctive mark in the world," by which, I take it, he is, at least in part, making the point we made earlier that consequentialism must say something about right decision. On the other hand, suppose that consequentialism is expressed as a doctrine about how to go about making the morally right decision, as a variety of subjective consequentialism in Railton's terms, and suppose in particular that it says to think along φ lines. What then if thinking along φ lines is discovered to have bad consequences in certain situations? (Jackson 1991, 470)

Jackson proposes a new account of right action that is neither fully objective nor fully subjective. In this account, neither outcome nor motivation is necessary to determine the rightness of an action, though they both are sufficient conditions. He states that "What is true is that doing an act for the right reason is sufficient but not necessary for it being what ought to be done in the sense we are insisting is central in ethics" (Jackson 1991, f.n. 20). If one makes a justified decision, a decision based on right motivation, regardless of the outcome, then one has done the right thing. Also, if one makes a decision that leads to the best outcome but was based on evil motivation, then the decision is still right. In a footnote, Jackson states that "on my view, consequentialism does not imply that a morally good intention is essential to a morally good act, at least if morally good act here means what an agent ought to do. It is possible to the right thing for the wrong reason. For an act which maximizes expected moral utility, and it might be that which prompts the agent to action" (Jackson 1991, f.n. 20).

Jackson proposes a theory of right motivation that is based on a combination of right desires and beliefs. Desires are idealized to and defined by a value function that is determined by the objective ranking of the possible outcomes based on their objective utility. According to Jackson, desires "rank the states of affairs in terms of how much the person would like the state of affairs to happen" (Jackson 1991, 464). Beliefs are the subjective probability function, or the degree of credence, that the agent assigns to each

outcome. This function indicates what an agent in fact believes, instead of what an agent ought to believe. Hence, in Jackson's account, a morally right action is an action that either has the best outcome or is chosen by the right motivation that was also partially determined by the best outcome. This partial determination is due to the contribution of desires to the criterion of right motivation.

Encountering motives in the account of rightness enables Jackson to solve the dilemma; it solves the prescription problem without making consequentialism self-defeating. Jackson argues that the "decision-theoretic account of consequentialism disarms the second horn of the dilemma by answering that in such situations the agent ought not to think along φ lines, for the agent's beliefs will then include that thinking along φ lines, in such situations has low expected moral utility" (Jackson 1991, 470). What an agent ought to do is to have desires that rank the alternatives in accordance with the consequentialist value function. Then, the subjective probability function, which is the idealized and "quantitative guise" of the agent's beliefs, multiplied by the value function, tells the agent what the right decision is. In this process, the agent is maximizing expected utility without consciously aiming for it.

Decision-theoretic consequentialism disarms the second horn of the dilemma by rejecting commitment to the view that maximizing expected moral utility is the right motive for action. The consequentialist value function to which the agent's desires should conform does not assign any additional value to what maximized expected utility since that would be "double counting" (Jackson 1991, 471). The subjective probability function, which indicates the agent's beliefs, does not have anything to do with maximizing expected utility either. So, maximizing expected utility cannot be the motive for action. Thus, consequentialism prescribes the right motive for action that guarantees the right decision based on the objective outcomes, regardless of whether *in fact* the action has the best outcomes or not, and without demanding that the agents have the motivation of maximizing utility or expected utility.

Jackson's proposal responds to the objection that Michael Stocker (1976) raises for modern ethical views. According to Stocker, considering motivation in the account of rightness of action implies that "a morally good intention is an intention to do the act for the sake of its goodness or rightness" (Jackson 1991, 469). However, according to DTC, "What ought to move a person to action according to consequentialism are desires which may be represented as ranking states of affairs in the consequentialist way, but maximizing expected utility is not a factor in this ranking" (Jackson 1991, 471).

In sum, the rule for action in DTC is to maximize *expected* moral utility instead of moral utility. The shift from utility to expected utility enables this theory to talk about

right actions in terms of right motivations, which leads to a criterion for right decisions. An action with right motivations can be right even if it does not have the best outcomes. This move obviates any need for commitment to any mental process as long as either the action maximizes utility or the decision is justified. In Jackson's term, DTC has "built into its very account of right action, a doctrine about right motivation" that "is not committed to any particular view about the mental process that an agent ought to go through in deciding what to do" (Jackson 1991, 468).

An action can be consequentially motivated without any need for consequentialist deliberation. Even when the action does not maximize utility, the agent's decision is right if it is justifiable. But the justification does not require any mental process. In Jackson's terms, "sometimes you ought not to go through any mental process at all" (Jackson 1991, 472). For example, Jill has to decide between drug A and B, and she spontaneously chooses drug A without even thinking about it consciously. But it turns out that drug B was the better choice. In this case, Jill's decision is justified since she knew that it is more likely for drug A to improve the symptoms than drug B. Jackson supports this justification by arguing that "spontaneous action is not action without belief, it is action without conscious reviewing of belief" (Jackson 1991, 472). In sum, if, without any conscious consequential deliberation, Jill's desires are aligned with the consequential value function, and she applies her beliefs to them rationally, then she is not blameworthy. To determine the justification for her action, however, it does not matter whether she actually applied the beliefs rationally or whether she knew what she desired. What matters is whether it is possible to describe her decision from a third-person perspective and to attribute the right desires to the application of her beliefs in that passive description.

Objections

The main advantage of DTC is the justification of a decision. The prescription problem can be interpreted in two different ways. One interpretation has to do with the way the criterion of rightness guides the agent to in fact achieve the desired outcome. The main concern in this interpretation is to prescribe a decision process that guarantees the best outcome; I call this interpretation the heuristic concern. The other interpretation has to do with how the criterion of rightness can prevent the agent from being blamed. The main concern of this interpretation is the justification of the prescribed action; I call this interpretation the justificatory concern. The justificatory concern is the key to connecting Jackson's prescription problem to the problem of moral luck. In what follows, I argue that DTC and OC both leave the heuristic concern of the prescription problem

to be resolved by empirical evidence. The main difference between DTC and OC is that DTC attempts to solve the justificatory element of the prescriptive problem while OC does not. However, I argue that DTC is not successful in its attempt since it raises other problems that I discuss shortly.

In his paper, "Alienation Consequentialism and the Demands of Morality," Peter Railton addresses Jackson's prescription problem in its heuristic sense. He considers the objection that the "lack of any direct link between objective consequences and a particular mode of decision making leaves the view too vague to provide adequate guidance in practice" (Railton 1984, 116). Railton's solution for this problem, which leaves Jackson unsatisfied, is that "objective consequentialism sets a definite and distinctive criterion of right action, and it becomes an empirical question… which modes of decision making should be employed and when" (Railton 1984, 116). Railton's response, similar to DTC, leaves the decision process a matter of empirical question. OC and DTC agree that the decision process that in fact maximizes utility or expected utility is highly context dependent and should be a matter of empirical evidence. If lack of deliberation is the method that achieves the desired outcomes, neither OC nor DTC demands conscious thinking.

Rationality, in its narrow sense, is the main idea behind the criterion for justification of a decision for DTC. DTC ends up with a set of objective requirements that if an agent satisfies, she will not be blameworthy. These requirements are mainly derived from the idea that the agent's decision should be rational. Thus, regardless of how a person in fact came to a decision, as long as it is possible to interpret her decision process in terms of maximizing expected utility and as rational, the decision is justified and not blameworthy. As discussed in the previous section, DTC is not committed to any specific mental process; indeed, it does not require any sort of mental deliberation to make an action justified. This lack of commitment to any mental process is possible because the justification criterion is not based on what the agent in fact does but rather on an after-the-fact description of the action.

The rationality criterion proposed by DTC does not accommodate the justificatory concern. In other words, it is possible to act in accordance with the criterion for right actions in DTC and still be morally blameworthy. In example 1, I provide a case in which intuitively the agent seems blameworthy, and morally unjustified, while her decision seems rationally justified.

Example 1: Jill is about to leave work when she figures her expensive watch has been stolen. She knows that the watch was on her desk all day and that no one entered or left her office after she got there. Six people are in the room, and she does not know

any of them personally. However, one of the people in the room is African American. She knows that African Americans commit 60% of the total crimes in her country. Thus, if we take Jackson's criterion seriously, regardless of whether she is right or not, she is morally justified to believe that there is a high chance that the African American person in the room is to blame. However, the rational justification of this conclusion does not make it morally justified. There is a strong intuition[2] that she is in fact blameworthy if she thinks the person who belongs to a group that commits more crime in society has more likelihood of being the one who stole her watch.

Example 1 is one of many possible examples that show that rationality does not provide moral justification for action. Jill did not act in accordance with her conclusion, so the outcomes are irrelevant to her blameworthiness. She made a rational decision, and she has the right desires that want to find the criminal. Thus, her motivation is right. But she is still intuitively blameworthy. This problem can be traced back to the narrow definition of rationality in decision theory. The formal model of rationality that is introduced by decision theory is notorious for over-simplification and/or over-rationalization of human behavior (Sobel 1994).

Various modifications of the prisoner's dilemma are the standard counterexamples against the narrow definition of rationality. This famous example in game theory is used to show that a rational decision is not the best decision in terms of maximizing overall utility (Joyce 2007; Hitchcock 2016). The prisoner's dilemma is a standard case with usually two participants who each need to decide between two alternatives. However, the decision of the other participant partially determines the outcome of the participant who is deciding. If the agent acts in accordance with the decision-theoretic criterion of rationality, the outcome will be worse overall. A rational decision, in this case, is to "play safe" and not assume that the other agent will collaborate. However, this decision guarantees a worse outcome. The "irrational" decision is to assume the collaboration of the other agent, and it achieves the best outcome overall. In sum, if best consequences for everyone is concerned, in the prisoners' dilemma it seems justified for each individual to not act "rationally."

2. This intuition is due to her action falling into the category of discrimination, which "is prohibited by six of the core international human rights documents. The vast majority of the world's states have constitutional or statutory provisions outlawing discrimination (Osin and Porat 2005). And most philosophical, political, and legal discussions of discrimination proceed on the premise that discrimination is morally wrong and, in a wide range of cases, ought to be legally prohibited" (Altman, 2016).

DTC has an inevitable problem that I call the desire-luck problem. DTC pushes the moral-luck—and thereby prescription—problem one step back to the desires, which leads to problems similar to those that motivated its proposal. The formulation of the desire-luck problem is as follows. Our moral assessment seems appropriate for matters over which an agent has control. We do not have full control over our desires, but according to DTC, we are justified in assessing people morally based on their desires. Jackson argues that the desire that the agent ought to have is the one that conforms to the consequentialist value function. However, there is no prescription for how agents should acquire such desires, and the psychological studies that I discuss shortly suggest that it is not always possible to have full control over our desires.

The desire-luck problem causes a dilemma for DTC in the same way that standard consequentialism was subject to a dilemma. If DTC does not prescribe how we can get to the right desires, it needs to deal with cases in which the desires are affected unbeknown to the agent. Instances of implicit bias and situational bias, like the bystander effect, show how vulnerable are our desires to the effect of things that we do not have much control over. In fact, there might be no way for us to realize that those affects exist without professional help. On the other hand, if DTC assumes control over desires and prescribes a method, this could be self-defeating. It may turn out that such a method *in fact* leads to desires that will not conform to the consequentialist value function in other contexts.

We do not have full control over what affects our desires, so the fact that they conform to the consequentialist value function may be due to pure luck. Many studies suggest that our ranking of alternatives is affected by things that we are not aware of and that we do not have full access to how they affect our judgment. There are studies that suggest that our desire to help a person, or our judgment about how serious her situation is, might be affected by how much of a hurry we are in. Studies about the bystander effect also suggest that whether we rank helping a victim high among the alternatives is significantly affected by whether other people are willing to help the victim or not (Asch 1951; Carlson et al. 1988; Rodin 1969). The common feature of these studies is that their participants show significantly less desire to help in certain mental states or in a situational context compared to normal contexts and that they are unaware of this difference. However, none of these uncontrollable effects on desires prevents us from blaming someone who doesn't help a victim because she is late to work. Therefore, lack of access to what changes our desire makes having the right set of desires in a particular situation a matter of luck.

DTC can be self-defeating if it provides guidance for how to adopt the right desires. A person might do everything that maximizes expected utility, but in doing so,

unbeknown to her, her desires could be affected to the point that they do not conform to the consequentialist value function anymore. In what follows, I provide an example to make this point. Then I explain why DTC needs to talk about desires in an intuitive sense and why defining desires in terms of emotions seems like a natural option. I use an account of learning for emotions that seems more compatible with decision theory, and finally, I discuss why the provided example leads to a contradiction for DTC. The following example is a case in which desires are adjusted to maximize expected utility in one context, but the permanent change in desires makes the agent blameworthy in other contexts.

Example 2: Jalisa is a good nurse, but in order to do her job and stay sane, she has, over time, lost her sensitivity to people's pain. She does not prioritize someone's pain over answering a phone call or over something that she can do to help, for instance, stitching someone's wound and not be emotionally distressed by the patient's pain like a normal person. This manner is in fact motivated by all professionals in the community since it helps them avoid the effects of emotional distress in their decision-making process. However, her insensitivity to pain that came from years of working as a good nurse hurts her new partner's feelings. Jalisa's reaction to her partner's pain is far from what it needs to be.

Jackson's project can be summarized in three major moves, but for them to be valid, formal desires must be connected to an intuitive understanding of desires. His notion of desire is simply to formal and abstract to be what intuitively we expect desires to be. But, the plausible option for this connection is to use emotions to make the connection. In his first move, Jackson argues for the importance of motives for a moral theory and for justification of a decision. In the second move, he uses the intuition that motivation is composed of emotions and feelings. Finally, in his third move, he uses decision theory, which describes decisions in terms of a subjective element that he calls beliefs and an objective element that he calls desires. However, to transition from the second to third move, Jackson needs to show that his definitions of desires and beliefs are close to what intuitively composes motives. The formal notion of desires, common in linguistic and decision-theoretic descriptions of desire-like mental states, is usually understood as having a close relationship with emotions. The common view about the role of emotions among philosophers is that "Emotions make certain features of situations or arguments more prominent, giving them a weight in our experience that they would have lacked in the absence of emotion" (de Sousa 2014). This view about the role of emotions has important similarities with the role of the preference function in decision theory that

Jackson defines as desire. Moreover, other considerations about the nature and the role of emotions and desires makes an appeal to understanding desires in terms of emotions.[3]

Understanding desires in terms of emotions enables us to talk about how an agent can have control over what she desires. Emotions are learned by association with "paradigm scenarios." A type of a situation and a set of characteristics of expected responses to the situation are two elements of a "paradigm scenario" (de Sousa 2014). A complex and controversial mix of biological and cultural factors determines what the expected responses are in each type of situation (de Sousa 2014). The process of learning, however, happens through time by associating the proper responses to each paradigm scenario. For some more fundamental emotions, the associations are "drawn first from our daily life as small children and later reinforced by the stories, art, and culture to which we are exposed. Later still, they are supplemented and refined by literature and other art forms capable of expanding the range of one's imagination of ways to live" (de Sousa 2014).

Jalisa is blameworthy in the context of her relationship, but her blameworthiness is a natural consequence of following the prescriptions in another context. In Jalisa's example, the paradigm scenario is a type in which someone is in pain and needs help. In the context of the hospital that she works in, it is expected that she not feel any emotional distress and not let the patient's pain change her desires or her preference function. But in the context of her new relationship, Jalisa needs to be sensitive to her partner's pain. The right desire in this context for Jalisa is to prioritize her partner's pain. But when Jalisa successfully learns to have the appropriate desire in one paradigm scenario, it is not possible for her to immediately have the right desire in a novel context, namely in her relationship, that triggered the same paradigm. Jalisa's motivation and therefore her decision in response to her partner is wrong and blameworthy since she does not have the right desire or the right preference function. Still, her blameworthiness for her

3. The following is de Sousa's explanation for why this account of emotions makes it appealing to understand desires in the way that Jackson talks about in terms of emotions: "This account does not identify emotions with judgments or desires, but it does explain why cognitivist theorists have been tempted to make this identification. Emotions set the agenda for beliefs and desires: one might say that they ask the questions that judgment answers with beliefs and evaluate the prospects that may or may not arouse desire. As every committee chairperson knows, questions have much to do with the determination of answers: the rest can be left up to the facts. In this way emotions could be said to be judgments, in the sense that they are what we see the world 'in terms of.' But they need not consist in articulated propositions. Much the same reasons motivate their assimilation to desire. As long as we presuppose some basic or preexisting desires, the directive power of 'motivation' belongs to what controls attention, salience, and inference strategies preferred" (de Sousa, 2014).

decision and her desire is due to the DTC prescription for right motivation and to her normal response to learning the right desire in a paradigm scenario.

In sum, Jackson uses the DTC to solve the prescription problem as the main advantage of this theory over OC. However, a normative theory needs to address our moral intuitions. I chose an example of racial discrimination to motivate the intuition that moral reasoning might not be easily captured by a general claim about how we should/we do make decisions. The details about the context matters in a way that is not easily captured by decision theory. Nurses are the paradigm example of a morally good person with consequentially praiseworthy contributions. I used an example of a nurse to show that right desires cannot make up for our expectations of right motivation. If being a nurse is a good, and being a better nurse is even better, it is praiseworthy that Jill wants to be a better nurse. But DTC considers a nurse blameworthy in some contexts although it prescribes being a good nurse.

References

Asch, Solomon. 1951. *Effects of Group Pressure upon the Modification and Distortion of Judgment.* Pittsburg: Carnegie Press.

Altman, Andrew. 2016. "Discrimination." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta.

Ben Ze'ev, Aaron, and Ruhama Goussinsky. 2008. *In the Name of Love: Romantic Ideology and Its Victims*. Oxford: Oxford University Press.

Carlson, Michael, Ventura Charlin, and Norman Miller. 1988. "Positive Mood and Helping Behavior." *Journal of Personality and Social Psychology* 55 (2): 211–29.

de Sousa, Ronald. 2014. "Emotion." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta (ed.).

Illouz, Eva. 2012. *Why Love Hurts*. Cambridge: Polity.

Jackson, Frank. 1991. "Decision-Theoretic Consequentialism and the Nearest and Dearest Objection." *Ethics* 101 (3): 461–82.

Nagel, Thomas. 1979. *Mortal Questions*. Vol. 89. Cambridge University Press.

Osin, Nina, and Dina Porat, eds. 2005. *Legislating Against Discrimination: An International Survey of Anti-Discrimination Norms*. Leiden: Martinus Nijhoff.

Railton, Peter. 1984. "Alienation, Consequentialism, and the Demands of Morality." *Philosophy and Public Affairs* 13 (2): 134–71.

Rodin, Judith, and Bibb Latané. 1969. "A Lady in Distress: Inhibiting Effects of Friends and Strangers of Bystander Intervention." *Journal of Experimental Social Psychology*: 189–202.

Stocker, Michael. 1976. "The Schizophrenia of Modern Ethical Theories." *Journal of Philosophy* 73 (14): 453–66.

Journal of Cognition and Neuroethics

Grammar is NOT a Computer of the Human Mind/Brain

Prakash Mondal
Indian Institute of Technology Hyderabad

Biography

Prakash Mondal is Assistant Professor of Linguistics and Cognitive Science at the Indian Institute of Technology Hyderabad, India, and the author of *Language, Mind and Computation* (Palgrave/Springer Nature, 2014), *Language and Cognitive Structures of Emotion* (Palgrave/Springer Nature, 2016), and *Natural Language and Possible Minds* (Brill, 2017).

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). February, 2018. Volume 5, Issue 1.

Citation

Mondal, Prakash. 2018. "Grammar is NOT a Computer of the Human Mind/Brain." *Journal of Cognition and Neuroethics* 5 (1): 85–100.

Grammar is NOT a Computer of the Human Mind/Brain

Prakash Mondal

Abstract

This paper will attempt to debunk the idea that human language grammar as part of the Faculty of Language (FoL) is intrinsically a computing device. The central argument here is that grammar does not compute. One way of demonstrating this is to show that the operations of grammar in the Generative model do not have the character typical of computations. Thus, the central operation of grammar Merge, which combines lexical items to produce larger expressions, can be defined as a recursive function, but it does not share the inductive properties of recursive functions in mathematics in view of the consideration that recursive functions define computability. On the other hand, if the language faculty is a computing system, the language faculty must inherit the halting problem as well. It is easy to impose the halting problem on the selection of lexical items from the lexicon in such a manner that FoL may or may not terminate over the selection of lexical items. We can say: there is no FoL way of telling if FoL will ever terminate on *x* or not when *x* is a selection from the lexicon. The halting problem for FoL is disastrous for the view that grammar is a computing system of the brain/mind since it detracts from the deterministic character of FoL. This has significant repercussions not just for grammar that cannot be restricted to any limited view of mental computation but also for the nature of the cognitive system as a whole since any cognitive domain that is (supposed to be) language-like cannot be said to compute as well.

Keywords

Grammar, Faculty of Language, Computation, Merge, Halting Problem, Cognition

Introduction

The idea that the grammar of human language is a formal system which forms part of the human mind realized as the Faculty of Language (FoL), and is a computing device has been advanced and popularized by the influential paradigm of Generative Grammar. The reason why this has been highly influential is that human language grammar is thought to be a kind of software installed on the brain hardware when a language is acquired by members of *Homo Sapiens*. Human language grammar, as part of our mind, which seems to be a non-physical intangible entity, is positioned at a higher-level of abstraction than the brain structures in which it is ultimately realized, just like the level of software in any modern day computer is situated at a level of abstraction removed from the level of the electrical circuitry. This makes the abstractness of the system of grammar

possible. Thus, this is supposed to explain why grammar as a formal system, just like a mathematical system (for instance, the system of arithmetic), generates an infinite number of linguistic expressions from a finite stock of words. While many theoretical linguists and cognitive scientists have, over the period, criticized and contested another influential view associated with the theory of Generative Grammar (Chomsky 1995, 2000), which is that humans are born with a biologically-hardwired capacity for language called Universal Grammar. Theoretical linguists and cognitive scientists have not found anything wrong with the idea that the language faculty, which is supposed to be a mental system, is *intrinsically* a computing device.

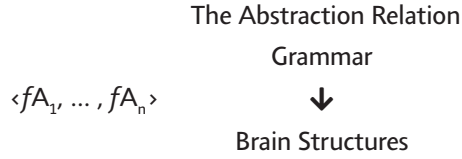


Figure 1: The Abstraction Relation

Here, grammar can be thought to abstract away from the underlying brain structures by way of a number of abstraction functions such as $fA_1 \ldots fA_n$, which perform a series of mappings to move away from the neurobiological structures and then reach into the higher level of abstraction from the lower-level of realization in the brain hardware. This ensures that grammar remains at a distance sufficient for a level of abstraction removed from the level of physical realization. Thus, for example, if there is an abstraction function fA_i that maps some neural state onto some state of the FoL, the idea is that if some state of the FoL changes to another state, then the designated change must be caused in some way by a change in the underlying brain state. There must be many such abstraction functions, since states of the FoL are abstractions whose configurations and transitions are to be related to the state changes in the neural hardware in ways that can be formally characterized. In this way, the state transitions in the physical system of the brain are to be mapped onto the state transitions of the FoL. This ensures that grammar, as part of the FoL, can execute operations that are sufficiently abstract, on the one hand, and are yet physically realized in the brain structures on the other. But, these operations are customarily understood and claimed to be computations, and grammar as a system performing such computations is thus a computing device. This is so because the FoL is supposed to be a computational system. Hence, before we proceed further, we need to

understand what computation is and, in what sense, grammar, or for that matter, the FoL is said to be a computational system.

In what Sense is Grammar (FoL) a Computing Device?

When one raises a question about whether something is computational, a lot depends on whether one deploys the right concept of computation to a phenomenon at hand, while checking if the given phenomenon falls under computation. The concerns are similar when we focus on language and wonder, in what sense, the system of grammar performs computations. Even though there is an amount of vagueness embracing the notion of computation applying to the FoL, it appears that linguistic computation, as performed by the FoL, jells well with the classical view of computation on which symbolic inputs are mapped to symbolic outputs according to some well-defined rules that are defined over those symbolic units which exist in the form of linguistic strings. This notion of computation lies in the narrowest region in the whole hierarchy of varieties of digital computation (Piccinini and Scarantino 2011). This notion of linguistic computation has been adopted in much of formal linguistics implicitly or explicitly because the representational vehicles of language are discrete in form. But, it is not clear whether we can take linguistic computation to be a kind of *generic* computation covering both digital and analog computations. Although this question may not a have clear answer, as most linguistic frameworks that apply the notion of computation do not demarcate the notion of computation, the answer is more likely to be no. What is noteworthy here is that the digital notion of computation has been the central principle in the model of cognition advanced in the field of cognitive science, in general, and presupposed in much of theoretical linguistics, in particular. The analog sense of computation is not in harmony with the view of linguistic computation on the grounds that analog computational processes are driven and determined by the *intrinsic* representational content of the vehicles that are analog in nature, whereas digital computation involves mappings of inputs onto outputs that are executed without any regard to the content-determining properties of the representational digital vehicles (O'Brien and Opie 2011). This squares well with what Chomsky (1980) has espoused, especially when he thinks of linguistic computation in the sense that the mentally instantiated system that can be identified with the FoL is a computing device of some sort. Thus, he maintains that '… "autonomous" principles of mental computation do not reflect in any simple way the properties of the phonetic or semantic "substance" or contingencies of language use.' From this, we can draw the conclusion that linguistic computation does not plausibly

encompass the analog sense of computation. Taken in the light of these considerations, linguistic computation cannot be regarded as a type of *generic* computation that embraces both digital and analog computation.

As we proceed, we shall require the appropriate notion of computation to accompany the critique that follows. This will help get a grip on the concept of linguistic computation when dealing with the view that grammar is a computing device of some sort. Central to any notion of computation are the following important elements: (i) a function that is computed; (ii) a system which computes the function; and, (iii) an effective procedure (also called an *algorithm*). This is also evident in the *Church-Turing Thesis*, which states that anything that can be computed with an effective procedure in the physical world can be computed in Turing machines. Related to this is the computational thesis in mainstream cognitive science that postulates that the human mind, or the brain, is itself a computing system. The rationale seems to make sense only when we suppose that the human brain is ultimately a physical object of some kind that can run computations just like any other physical system that executes computations. We may now wonder what the right physical system is that runs, or is appropriate for, linguistic computations. If we follow the lines of thinking adopted in Chomsky (1995, 2000), then we can make sense of the physical system fit for linguistic computation. In fact, the physical system for running linguistic computations is a system or module of the brain dedicated to language. Thus, it is the language faculty in the brain/mind that computes because the language faculty is conceived of as the physical organ specific for language situated within the confinements of our brain. The language faculty computes because it is supposed to have a *computational procedure* that engages in all kinds of linguistic computation. The next essential ingredient of computation is a function, or rather a computable function. The domain of such functions in formal linguistics may well correspond to the domain of formal operations that apply to structures to make structural distinctions of linguistic representations, when linguistic structures are inserted, erased and thereby altered. In this sense, we can take linguistic computations of the FoL to be the operations of a version of the Turing machine that works on a potentially infinite tape and reads, writes, or erases symbols on the tape. In a nutshell, the functions that can fall under linguistic computation are those which subscribe to the formation, substitution, and deletion of phrases, sentences, or larger linguistic expressions. Significantly, this is the *process-oriented* aspect of computation implicit in the specification of the Turing machine. That the system of grammar can be taken to be executing computable functions, or rather algorithms, accords well with the *abstraction-oriented* aspect of computation, which consists in the specification of computable functions that can be implemented by

algorithms in a system. If the computational character of the FoL is considered under the cover of the abstraction-oriented aspect of computation, then all that really matters is the specification of computable functions to be implemented in the putative computational system of the language faculty. That the specification of computable functions that grammar, as a system, is supposed to execute is possible has been shown for the central operations of grammar in the Minimalist model of grammar, especially for Merge, which is understood to concatenate syntactic objects (see Mondal 2014). In this context, Foster's (1992) notion of an algorithm as a sequence of transitions between the states of a machine is handy enough. In particular, Merge combines two syntactic objects to form a single syntactic object (which is actually a set). Thus, for a sentence like 'John loves a car', we have Merge (a, car) = {a, car} and Merge (loves, Merge(a, car))= {loves, {a, car}} and then (John, Merge (loves, Merge(a, car)))= {John, {loves, {a, car}}}. Note that the formulation of Merge for the generation of the linguistic expression 'John loves a car' is recursively specified--more will be said on this below. An algorithmic representation of the operations of Merge for the phrase 'a car' can thus be schematized as (1), by following Foster.

(1) $[SO_1$: a SO_2: car L: Σ] → [SO: a, car L: Σ] → [SO: {a, car} L: Σ] →

[SO: Σ {a, car}]

Here, SO is a syntactic object and L denotes the label of an SO, and Σ is the actual value of L. Thus, each item on the left of the colon is the label, and the one on the right designates the value of that label. Each item enclosed within braces represents a 'snapshot' of the state of a computation, and the arrow represents a transition between one such state and another. Now that the operation Merge has been shown to have a computational character in defining computable functions that can be coded as algorithms, the core generative engine of the FoL can be said to run computations by virtue of containing computable functions defined over the symbols the system of grammar operates on. This ensures that the system of grammar is viewed as a computing device whose computational nature can be characterized in the standard terms of the execution of computable functions.

Why Grammar (or FoL) is not a Computing Device

We may now look into the reasons why grammar or the FoL cannot be considered to be a computing device. Since the system of grammar is a computing device in virtue of defining computable functions, it possesses the abstract capacity of generating infinite

linguistic expressions *just as* an abstract system of arithmetic generates infinite arithmetic expressions. The relevant mathematical property here is the recursive character of the operations the system of grammar instantiates. An analogy from mathematics can be drawn in order to demonstrate how grammar as a system is *intrinsically* computational. Take, for instance, a recursive function that increments the value of a given number by 1: f(n)= n+1 when *n* is a natural number. Thus, f(5)= 5+1= 6 when *n*=5, for example. This function is recursively defined, in the sense that the given function is specified in terms of each calculated value of its own output. Hence, this function can also be specified in a manner that involves the invocation of the same function. So, we can write f(n)= f(n-1) +1. It may be noted that an inductive definition forms an intrinsic part of the formulation of the incremental function here. This is because the inductive definition licenses the inference that the function can be specified in terms of each calculated value of its own output by way of an invocation of itself. As shown above, the generative mechanism of grammar has a recursive characterization in virtue of the fact that the generation of an infinite number of linguistic expressions is part of the recursive definition of the operation Merge. That is, the putative computational system of the language faculty possesses this mechanism by virtue of having the operation called Merge. Therefore, it seems clear that all that matters is the specification of the function concerned, not how this function is implemented in the language faculty *in real time*. This must be so because Merge is defined as a function *in intension*. If grammar is a computing system in this sense (as far as the mapping function so defined is concerned), it is not unreasonable to think that the relevant properties of recursive functions that hold true for the set of natural numbers should also be found in the set of natural language expressions generated by Merge or by any conceivably analogous computational mechanism of grammar. Let's see how we can test this formal parallelism. Suppose we have the following sentences which are output by Merge:

(2) (Amy + (trusts+ (a + man + … + … +…)))

(3) (Amy +…+ … + (trusts+ (a + man)))

The sentence in (2) can be taken to have an unbounded expansion which goes on like this: 'Amy trusts a man who is known to have three mansions which are located in three different countries that form a certain contour around a place that defies any description …'. Likewise, (3) can be also be unboundedly long such that its expansion may run like: 'Amy who is one of the finest scholars at our university which motivates the study of culture in unexplored territories which may not have any access to education… trusts a

man'. The problem for Merge is that it cannot get off the ground in (2) since Merge, as a constructive operation, starts and continues to work in a bottom-up fashion, whereas it can never terminate, even if it does start in (3). Note that recursively defined functions in mathematics are such that they may never terminate, hence this particular argument certainly cannot have the appropriate force it should have, since, after all, Merge taken as a mapping function is defined in intension. However, functions operating on (the set of) natural numbers, as in $f(n)= n+1$, *at least* get off the ground when there are inputs to be mapped onto outputs, regardless of whether they terminate. To put it in other words, functions operating on (the set of) natural numbers do not spell out the problem of not starting in the first place, while Merge contains the germ of the problem of not starting in the first place as well as inheriting the problem of non-termination. One may try to circumvent this problem for Merge by postulating null items that are *assumed* to exist in the unboundedly long sentence in (2) in order to save Merge from getting stuck into this trap. If the same strategy is adopted, then the null items which may be taken to be the stand-ins or proxies for the relative clauses constituting the expansion in (3) may also be assumed to have been Merged. For all its appeal, this strategy is groundless because items empty of substance are inserted in a linguistic expression which is not even a well-formed expression and perhaps does not even exist due to its unbounded or unfinished form. We end up inserting items empty of substance into an expression, which, as a whole, is already empty of content. The result is anything but a meaningful statement. Plus, this detracts from the operational character of Merge because Merge does not concatenate null items. This is the case in the Minimalist model of the language faculty, for there is a ban imposed on the FoL that disallows items which have not present in the selected set of lexical items on which computations are to operate. Besides, null items for chunks as big as relative clauses cannot be selected from the lexicon, nor can they be justified on linguistic grounds, since nothing would then prevent one from postulating null sentences whether simple or complex.

The worry does not, of course, stop here. There is another, deeper, more fundamental problem residing in the postulation of formal parallels between recursive functions in mathematics and the putative computational mechanism of grammar. Just for instance, the principle of *mathematical induction* applies to all well-formed functions when it is used as a proof technique to test whether something holds for an infinite set because we cannot check all items in a potentially infinite set. So, as per the principle of mathematical induction, if some proposition P holds for n, it also holds for $n + 1$. The second step in this formulation constitutes an *inductive* generalization that may also be aligned with various other kinds of generalizations drawn inductively by human beings. Let's now reconsider

the example in (2) to determine whether mathematical induction can be applied to it. The example (2) has been represented the following way in (4).

(4) 'Amy trusts a man' + Rcl^k (where $Rcl^k = k$ number of relative clauses)

Since it is necessary to render (2) in a manner that makes it amenable to the application of mathematical induction, the formulation of (2) in (4) serves to demarcate the domain, that is, the portion Rcl^k, over which mathematical induction can be taken to apply. One of way accomplishing this is the following way of characterizing the relevant set so that we state that mathematical induction applies over the set in (5).

(5) {'Amy trusts a man', 'Amy trusts a man who is known to have three mansions',

'Amy trusts a man who is known to have three mansions which are located in three different countries' …}

But what are the appropriate properties of this set, or of the members of this set, that can help establish that some proposition precisely formulated holds for the *n+1*th expression only if it holds for the *n*th expression? In what sense can the expression 'Amy trusts a man' be supposed to be the *n*th expression? Or, in what sense can the expression 'Amy trusts a man who is known to have three mansions' be the *n+1*th expression and so on? What are the exact properties of these expressions such that their succession can mimic that of natural numbers when the natural numbers that are inputs or outputs of a function are defined in terms of a function? One suggestion that can be implemented here is that the relevant proposition that needs to be tested has to be formulated by tracking the *depth* of concatenation of relative clauses. That is, one may say that 'Amy trusts a man' is an expression with the value of the depth of concatenation fixed at 0, and similarly, 'Amy trusts a man who is known to have three mansions' has the depth of concatenation set at 1 and so on. This may be supposed to reflect the progression of these expressions at par with that of natural numbers. So the proposition to be tested is that the concatenation of a relative clause to a sentence whose verb phrase is transitive returns a well-formed expression of English. This can be couched in terms that may be supposed to ride especially on the inductive generalization that the attachment of a relative clause to a sentence whose verb phrase is transitive *always* yields a well-formed expression of English. The specific rule may be formulated in terms of *phrase-structure rules* familiar in formal linguistics. The advance of formulating such a rule is necessitated by the consideration that the rule has to be maximally general so that inductive definitions hold

true for (4) or even (5). Let formulation in (6) be the exact phrase-structure rule that we need to capture this inductive definition:

(6) Sentence (S) → Noun Phrase (NP) Transitive Verb Phrase (TVP) + Rcl^{k}

But the problem is that the rule in (6) can never ground (4) or (5) in an inductive generalization, simply because rules like this overgenerate. Nothing stops (6) from generating (7), below:

(7) *Amy trusts a man which is known to have three mansions which are located in oneself that forms a certain contour around hers that has any description …

Likewise, there is nothing that can prevent one from having an expansion in (4) at some *n+1* level that renders the whole expression grammatically illegitimate, as in (8), below:

(8) *Amy trusts a man who is known to have three mansions which will sold tomorrow.

An attempt to import grammatically relevant, context-sensitive information, and selectional properties of predicates and other expressions into the contexts of (7-8) will inevitably vitiate the prospect of having a rule that will possess such a general character as to be amenable to an inductive definition. This is because when we say that if the incremental property of the function *f*(n)= n+1 considered above holds for the number *n*, induction guarantees that it will also hold true of the *n+1*th number. That is, mathematical induction ensures and safeguards the *generality* of the induction without any provisos or conditions fixed for the induction to apply in the first place. Needless to say, this is doomed to fail for natural language expressions. There is the following dilemma when we turn to natural language. On the one hand, we require something like a function that can have the desired formal generality across a potentially infinite range of expressions, and on the other hand, the nature of natural language grammar is such that it defies the formulation of any such function. It is important to recognize, in this connection, that neither the compositional function nor the intuitive sense of concatenation can serve this purpose. The former is of no substantive value *in this particular case* because natural language abounds in non-compositionally formed expressions (idioms, for example). Thus, we can have expressions such as 'take for granted,' 'beat around the bush,' 'call time on,' etc. whose meanings are not strictly determined by the combinations of the meanings of the parts of the whole expressions. Concatenation, on the other hand, as

an operation is too trivial to have any linguistic value since the output expressions from the operation of concatenation can be deviant or ungrammatical. There is nothing that can, for instance, prevent one from concatenating 'an' with 'ball' or even 'for' with 'done', which will yield 'an ball' or 'for done,' both of which are ill-formed in English.

Beyond that, Merge cannot defined as a recursive function, given that recursive functions define computability. Thus, for example, addition is a recursive function because it can invoke itself as an input. So +(3, 5)=8 and then +(8, 8)=16 can be better expressed as +(+(3,5), 8)=16. Also, note that the inputs and outputs are all members of the set of natural numbers. This is not so for natural language. If Merge(John, runs)= John runs, we cannot have something like Merge(Merge(John, runs), John runs)=John runs John runs. That is, 'John runs John runs' is not a well-formed string in English. This can be generalized to any language other than English. The relevant property is called the closure property of functions or operations defined on natural numbers. Closure properties make it possible for natural numbers to be defined within the bounds delimited by the set of natural numbers. That is, it is closure properties of natural numbers that tell us that both 5 and 4 in 5+4 =9 are natural numbers and so is the number 9. Similarly, both the input numbers and the output number involved in the operation of multiplication in 5×7=35 are natural numbers. There is nothing in natural language that is even remotely closer to this mathematical property when we look at the relevant linguistic expressions. Therefore, the following expression in (9), which results from the Merging of 'Amy trusts a man' with 'Amy trusts a man who is known to have three mansions' is ungrammatical:

(9) * Amy trusts a man Amy trusts a man who is known to have three mansions.

Finally, and most importantly, it may also be supposed that the problem of non-termination is in general true of procedures specified in abstraction, given that all procedures in practical reality must terminate, and if so, the problem of non-termination cannot be characterized as a problem for the computational mechanism of grammar. As we shall soon see, this may not be a problem for mathematical functions, or even for the Turing machine, since they are intrinsically mathematical, or purely abstract objects, not anchored in any physical system, though they can be implemented or instantiated in a physical system. But, this does not hold true for the computational mechanism of the language faculty since the language faculty is by its *intrinsic* character a mental system or a mental organization. In fact, the *halting problem* (Turing 1936) that is intrinsic to the model of computation inherent in the specification of the Turing machine must also apply to the putative computational system of the language faculty if the mapping

function of the putative computational system of the language faculty is translated into the operations implicit in the specification of the Turing machine. Even if there could be *intensional* differences between the model of computation implicit in the specification of the Turing machine and the mapping function in standard mathematical formalisms of computability despite the fact that they are descriptively or extensionally equivalent (see Soare 1996), such intensional differences—whatever they turn out to be—cannot be brought forward in order to dodge the halting problem for Turing machines. The reason is that the extensional equivalence between the model of computation implicit in the specification of the Turing machine and the mapping function in formalisms of computability is all that matters to the extrapolation of the halting problem to the putative computational system of the language faculty. Any intensional differences arise from a certain way in which computations are looked at or viewed by humans, and this cannot be built into the language faculty itself. Nor can these differences ground a different *mode* of computational operations that avoids the halting problem, because the problem of non-termination inherent in the halting problem is a fundamental part of any formulation of computation abstracting away from the real world.

One way of demonstrating the problem is to take lessons from the *halting problem* (Turing 1936) that is intrinsic to the model of computation inherent in the specification of the Turing machine. If the language faculty is a computing system, the putative computational system of the language faculty must also face the vagaries of the halting problem. The language faculty in the Minimalist model of Generative Grammar (Chomsky 1995) selects lexical items from the lexicon and then applies the binary operation Merge that combines these lexical items, and finally maps the constructed objects to the sound system (Phonological Form) and the meaning system (Logical Form). It is easy to impose the halting problem on the selection of lexical items from the lexicon in such a manner that the putative computational system of the language faculty may or may not terminate over the selection of lexical items. By following Partee, Ter Meulen, and Wall (1990), we can define the halting problem for the language faculty the following way.

L={*x*: a TM accepts *x*}

Here, L is a language that is defined as a set of *x*s, which are the strings generated/accepted by a Turing machine TM. Let's assume that the *x*s here are discrete lexical items that can be drawn from the lexicon. When FoL is said to terminate on *x*, it actually completes the task of selection of *x* from the lexicon. So, we can have N={*x*: FoL selects and terminates over *x*}, where N is the set of *x*s, and this set can otherwise be conceived of as a list. The halting problem for FoL is simply this: there is no FoL way of telling if

FoL will ever terminate on *x* or not. Suppose that one insists that it is possible. We then have N′={E(FoL): FoL selects and terminates over E(FoL)} where E(FoL) is the encoding of an FoL that is the input to the same FoL itself. Since E(FoL) itself can be in the form of lexical items, just as *x* is, it is okay to have E(FoL) as an input. Since FoL is a computing system, N′ is decidable by FoL. From this, it follows that the complement of N′, which is, say, N′′, is also decidable by some version of FoL, say, FoL′. If so, we have N′′={x′: x′ not the encoding of any FoL, or else x′ is the encoding of an FoL that does not select and terminate over E(FoL). Now we can ask if E(FoL′) is a member of N′′. That is, is E(FoL′)∈ N′′.

If we assume E(FoL′)∉ N′′, then E(FoL′) is not one of the items selected by FoL′, and hence FoL′ does not select E(FoL′). Therefore, FoL′ is a version of FoL that does not accept its own encoding. This ends up making E(FoL′)∈ N′′ true. Contradiction! If, on the other hand, we assume that E(FoL′)∈ N′′, then FoL′ selects and terminates over E(FoL′). But, since E(FoL′)∈ N′′, by our assumption, FoL′ cannot select and terminate over E(FoL′). Hence, FoL′ does not select and terminate over E(FoL′). Again, a contradiction!

The halting problem for the FoL can prove to be fatal because the FoL may never execute computations, as computations get off the ground only when lexical items are combined through Merge. Additionally, this damages the deterministic character of the FoL on the grounds that the computational system that the FoL is supposed to be always maps the outputs of Merge to the sound and meaning systems without fail. The question of the FoL having options in its trajectories or paths of operations does not even arise, also because the computational system operates beyond space and time. Simply speaking, the system of grammar, as a computing device, does not cease to always map syntactic objects to semantic and phonological representations. If this is how computations are always supposed to operate over the symbols that syntactic objects are, then the FoL inherits a deterministic character from the way the linguistic machine functions. But, the halting problem imposed over the selection of lexical items, which actually kick-starts the computational processes in the FoL, undermines this determinism, or rather, this deterministic mode of operation of the FoL. This is guaranteed by the fact that now the FoL may sometimes execute computations or may not terminate on the already started computational processes. This is *not* how the FoL is thought to function since the FoL is assumed to offer an optimal solution to the demands posed by the interface systems (the conceptual and articulatory-perceptual interfaces of the brain/mind) connected to the meaning and sound systems. That is, if the FoL is supposed to ensure a kind of optimization between the computational processes of the syntactic engine and the demands placed on it by the interface systems of the mind, the halting

problem for the FoL eliminates the possibility of the FoL offering an optimal solution to the demands posed by the interface systems. If there is no lexical selection, then there is no computation that can run over the symbols picked up from the lexical inventory. If this happens in some non-deterministic way, then the FoL *may* or *may not* meet the requirements placed by the interface systems, thereby turning into an inelegant system, though it is supposed to be an elegant system of optimization.

Implications

The upshot of the whole discussion in this paper is that grammar does not compute. If grammar does not compute, then there is no reason to think that this conclusion will block humans from producing and comprehending an unbounded number of linguistic expressions. The question of whether and how humans produce and comprehend an unlimited number of linguistic expressions has nothing whatsoever to do with the computational character of the device itself. Let's take the following consideration. While languages may or may not be infinite, in the appropriate mathematical sense of the term, this has nothing to do with whether humans can or cannot produce and comprehend an unlimited number of linguistic expressions (see Lobina 2017). Here, the symbolic nature and form of languages is dissociated from the psychological capacity of humans to make use of the symbolic system of language to produce/comprehend an unbounded number of linguistic expressions. In a similar vein, if grammar, as part of the FoL, which is a mental system, does not compute, then this does not in any way block the possibility that humans can produce and comprehend an unlimited number of linguistic expressions. The reason for this is that grammar, when conceived of as a computing device within the Minimalist model of the language faculty, is taken to be frozen from real space and time constraints, this makes grammar more like a purely abstract symbolic system devoid of contact with the real world. But, the fact that humans produce and comprehend an unlimited number of linguistic expressions has some connection to the real space and time considerations, as humans do process linguistic expressions of an unbounded number *in real time and space*. Therefore, even if grammar as a purely abstract system does not compute, this, in itself, has nothing to allow or disallow the human psychological capacity to produce and comprehend an unlimited number of linguistic expressions. This licenses the conclusion that the psychological capacity of humans to produce and comprehend an unlimited number of linguistic expressions *must* be ultimately segregated from the purely abstract properties of grammar, whether admitting of the conception of grammar as computing device in itself.

This has significant consequences and implications for cognitive computations within language, and other cognitive domains, such as vision, motor systems, etc. Any mental system or domain that is supposed to be language-like, or to possess properties and features (such as systematicity, compositionality, etc.) that are usually attributed to language cannot be said to be computing. Nor can it be regarded as a computing device of some sort. The cognitive system(s) responsible for thought and reasoning can be a good example here, since the system of thought has often been modeled on the system of grammar, as in the well-known Language of Thought (LoT) Hypothesis (Fodor 1975, 2001). Other cognitive domains, such as vision, motor systems, or memory cannot also be regarded as systems that run computations on the grounds that problems orthogonal to the halting problem imposed on lexical selection within the FoL can be extended to these systems. Thus, for example, the visual system can be said to run computations on the selection of properties or features such as size, color, shape, texture, depth, etc. of the perceptual world whose inputs are processed inside the visual system. Likewise, the motor system can be thought to operate on analog values of coordinates of body parts to run the appropriate computations for motor actions, and the memory system can be said to work on snapshots of events and things in order to assemble, de-assemble, sequence, retrieve, and erase memories. In all such cases, the halting problem imposed on the initial process of computations can scupper the computing device a cognitive domain, such as vision or the motor system, is imagined to be. This is more so because cognitive domains such as vision or memory often need to offer sub-optimal solutions to the problems posed by the messy world, and the mappings between designated inputs and outputs are not always straightforwardly driven by content-less properties of the symbols over which computations are supposed to run. That is why various effects, such as visual hallucinations and illusions, false memories, and memory blocking, etc. exist.

There are parallels between this work and the profoundly significant demonstration by the mathematical logician Kurt Gödel that mathematics as a formal system cannot be reduced to any delimited set of principles. Just as mathematics will always remain a never-depleting stream producing new theorems that cannot be bound within any predefined confinements of axioms, the formal system of grammar will always remain an ever-productive system generating newer and newer axioms and constraints of language which cannot be restricted by, and thus reduced to, any limited notion of mental computation.

References

Chomsky, Noam. 1980. *Rules and Representations*. New York: Columbia University Press

Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, Mass.: MIT Press.

Chomsky, Noam. 2000. *New Horizons in the Study of Language and Mind*. Cambridge, Mass.: MIT Press.

Fodor, Jerry. 1975. *The Language of Thought*. Cambridge, Mass.: Harvard University Press.

Fodor, Jerry. 2001. *The Mind does not Work that Way*. Cambridge, Mass.: MIT Press.

Foster, Carol Lynn. 1992. *Algorithms, Abstraction and Implementation: Levels of Detail in Cognitive Science*. San Diego: Academic Press.

Lobina, David. 2017. *Recursion: A Computational Investigation into the Representation and Processing of Language*. New York: Oxford University Press.

Mondal, Prakash. 2014. *Language, Mind and Computation*. London: Palgrave Macmillan.

O'Brien, Gerard, and Jon Opie. 2011. "Representation in Analog Computation". In *Knowledge and Representation*, edited by Albert Newen, Andreas Bartels and Eva-Maria Jung, 109-128. Stanford: CSLI Publications.

Partee, Barbara, Alice Geraldine Baltina Ter Meulen, and Robert Wall. 1990. *Mathematical Methods in Linguistics*. Heidelberg: Springer.

Piccinini, Gualtiero, and Andrea Scarantino. 2011. "Information processing, Computation and Cognition". *Journal of Biological Physics* 37:1-38.

Soare, Robert. 1996. "Computability and Recursion". *The Bulletin of Symbolic Logic* 2 (3): 284-321.

Turing, Alan. 1936. "On Computable Numbers with an Application to the Entscheidungs Problem". *Proceedings of the London Mathematical Society* 42 (2): 230-265.

Journal of Cognition and Neuroethics

Interoceptive Inference and Emotion in Music: Integrating the Neurofunctional 'Quartet Theory of Emotion' with Predictive Processing in Music-Related Emotional Experience

Shannon Proksch
The University of Edinburgh

Biography

Shannon Proksch has recently completed her master's degree at the University of Edinburgh. Most of her work centers on music and language cognition, especially the increasing role of music in empirical and theoretical studies of embodied, enactive, and social cognition.

Acknowledgments

This paper was written in satisfaction of the MSc in Mind, Language and Embodied Cognition at the University of Edinburgh. It was subsequently presented at the Mind and Brain conference in Flint, Michigan. Thanks are due to Jami Anderson and the conference organizers. Many thanks as well to my supervisor, Dr Alistair Isaac for his guidance and support; to my music professor Dr Nikki Moran and the Edinburgh Music Psychology Research Reading group for fostering my early ideas; and to my friends and family who have provided detailed feedback and encouragement throughout this project.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). February, 2018. Volume 5, Issue 1.

Citation

Proksch, Shannon. 2018. "Interoceptive Inference and Emotion in Music: Integrating the Neurofunctional 'Quartet Theory of Emotion' with Predictive Processing in Music-Related Emotional Experience." *Journal of Cognition and Neuroethics* 5 (1): 101–125.

Interoceptive Inference and Emotion in Music: Integrating the Neurofunctional 'Quartet Theory of Emotion' with Predictive Processing in Music-Related Emotional Experience

Shannon Proksch

Abstract

In this paper, I discuss how the predictive processing framework expands upon traditional, bottom-up, two-factor theories of emotion as passive physiological evidence-building and subsequent cognitive appraisal (James, 1884; Lange, 1885; Schachter & Singer 1962), by incorporating active inference based on predictive models of the causes of external and internal stimuli (Seth & Critchley 2013; Seth & Friston 2016). Accordingly, I posit that emotional content from music is evoked as a result of active exteroceptive inference related to the physical musical stimuli (based on statistical regularities of the current musical event and past experience) as well as active interoceptive inference regarding the listener's current autonomic, and physiological states. In addition, I propose that this general predictive processing framework is implemented through a 'quartet' of neurofunctional mechanisms (Koelsch et al 2015) which are dynamically implicated in the neural and physiological processes underlying general, and music-related emotional experience. Conceptualising emotion as active inference over both external and internal processes, implemented and maintained through a dynamically interacting subset of neural pathways as offered by the Quartet Theory of Emotion, provides a more detailed mechanism by which music evokes emotion and results in the subjective difference in the reported emotional experience of music between individuals.

Keywords

Music, Emotion, Predictive Processing, Active Inference, Quartet Theory of Emotion, Brainstem, Diencephalon, Hippocampus, Neurofunctional, Social Cognition

Part 1

Pre-Show

How and what emotional content is communicated or evoked by music constitutes a central question for music cognition. It is a question that is especially difficult to investigate, since the emotional response to a given piece of music can vary so widely in content and precision from person to person. One popular story in music cognition is that music-related emotions arise from the violation and confirmation of musical

expectancies based on statistical regularities of *external* musical features from an agent's past experience (Meyer 1956; Narmour 1990; Huron 2006). This fits well with predictive processing models of emotion, which are also based on expectancies, or predictions, that are honed by past experience. In predictive processing (PP) models, emotion arises from the process of minimizing prediction error[1] from active top-down predictions of the causes of *internal* bodily states (Hohwy 2013; Seth & Critchley 2013; Clark 2016). However, each of these accounts of emotion alone are too coarse grained to fully explain the underlying process of emotion in music and in general.

The predictive processing framework expands upon traditional, bottom-up, two-factor theories of emotion as passive physiological evidence-building and subsequent cognitive appraisal (James, 1884; Lange, 1885; Schachter & Singer 1962), by incorporating active inference based on predictive models of the causes of external *and* internal stimuli (Seth & Critchley 2013; Seth & Friston 2016). In the same vein, thorough descriptions of music-related emotional experience need to appeal to the *active* processing of expectancies of both interoceptive (internal, bodily) information as well as exteroceptive (external, structural) information, an explanation that is offered by the minimization of prediction error achieved through *active inference* within predictive processing.

However, relying heavily on the involvement of the insular cortex, with offhand mention of other emotional systems, these active inference accounts of emotion and interoception are importantly lacking the details of a distinct mechanism over which active inference is taking place. Such a mechanism can be described by appealing to neurofunctional models of emotional experience, (such as the Quartet Theory of Emotion (QTE) as outlined by Koelsch et al. 2015) which take into account how the brain and associated biological systems actually effect homeostatic maintenance and generate affective experience. I posit that incorporating QTE's multifaced neurofunctional mechanism can solve problems concerning the limitations within predictive processing accounts of emotion.

I propose that a network of interlinked neural systems, as offered by QTE, provides a set of low-level, high-level, and integrated neural mechanisms over which I claim PP occurs simultaneously, thus contextualizing the prediction error minimization within and between each system. Integrating this neurofunctional mechanism with predictive

1. Prediction Error Minimization (PEM) is basic principle underlying the predictive processing framework with the goal of "creat[ing] a closer fit between the predictions of sensory input based on an internal model of the world and the actual sensory input" (Hohwy 2013). This may be achieved in two ways: model-updating, or world-updating—as will be addressed further in this paper.

processing accounts of emotion in this way provides a more nuanced explanation of emotional experience than either theory offers alone—a claim that I will show is justified through an exploration of how QTE and PP, together, enhance our understanding of music-related emotions, shedding light on the answer to the question of how and what emotional content is communicated or evoked by music.

Set List

I will first provide a brief outline of current theories of music-related emotions and their proposed underlying mechanisms. Various mechanisms proposed by Juslin and Västfjäll (2008) will be shown to be linked by the further underlying component of expectancy. I will then explore predictive processing accounts of emotion, specifically active interoceptive inference, to address what theories of musical emotion and expectancy have right so far, as well as offer a further incorporation of interoceptive, physiological processes. Following this brief explanation of active interoceptive inference, a specific neurofunctional model as put forth by Koelsch et al (2015) will be reviewed: The Quartet Theory of Emotion (QTE). I will show that active interoceptive inference, in conjunction with the QTE, offers a mechanistically and functionally appealing account of emotional processes in general, and music-related emotions in particular. This project will end with a brief introduction to the explanatory value of combining interoceptive inference with QTE through reflecting on a series of vignettes of two fictional concert-goers with differing music-related emotional experiences.

Audience Introductions

> Music is the shorthand of emotion.
> —Leo Tolstoy

If this quote is accurate, then listening to music with someone might be a quick shortcut to sharing a similar emotional experience. We can listen to our country's national anthem at an Olympic games ceremony and share in the social emotions of pride and happiness or to a funeral hymn and share in the grief of our friends and family. However, it is easy to consider a case in which individuals do not share the same emotional experience when listening to the same music. Below, I introduce two fictional concert-

goers, whose story will help us to evaluate the explanatory upshot of an integrated QTE + PP model in the case of music-related emotional experience.

Audience Spotlight 1

Let's reflect on the summer that Rylan, a college student in her late teens or early twenties, takes Henry, her grandfather, to a Rammstein concert. Henry wants to know what the young kids are listening to these days, and Rylan promised him that he would love their music. In fact, these industrial metal, hard rock performers make up Rylan's favorite band. Although she doesn't know German, their concerts always leave Rylan feeling energized and exuberant, as if she is right where she belongs.

Henry prefers country music oldies: Merle Haggard, Patsy Cline. But he loves his granddaughter, so he had happily tagged along for the experience. This concert, however, leaves Henry feeling agitated and uneasy, as if he distinctly doesn't belong.

Rylan and Henry are listening to the same music, with the same exteroceptive information stimulating their senses, and surrounded by the same physical context in a concert environment. However, they seem to be having substantially different emotional responses. There is something occurring in each concert-goers body and brain besides merely processing and reaction to external musical stimuli.

Part 2

Current Attempts at Understanding Music-Related Emotions

The Search for Underlying Mechanisms

Much has been said about the relationship between music and emotion. Music, as a language of emotion, is meant to evoke an emotional response in its listeners and those who partake in musical experience. Many theories have attempted to account for this pivotal role of emotion in music, or music on emotions, from appeals to the extra-musical associations such as the context of a sad moment of a play, or appeals to the "choreographing of expectation" (Meyer 1956). Rylan might enjoy Rammstein's music because it reminds her of her trip to Germany. Henry might dislike the concert because the music reminds him of German-language videos from his primary school lessons on the horrors of the Holocaust. However, it has been noted that, despite the plethora of accounts of music-related emotions and varying appeals to extra-musical associations, the current trajectory lacks a focus or explanation of the underlying neural and physiological

mechanisms by which music effects emotional responses. Juslin and Västfjäll (2008, henceforth J&V) attempted to address this issue through identifying and explaining the underlying mechanisms behind music-related emotional experience.

These theories reviewed by J&V, as with most theories of emotion, are appraisal theories. They rest on the assumption that an emotion arises as a result of a cognitive appraisal, or a subjective evaluation informed by context on the personal level, relating to life goals or survival functions, and to which music-related emotions generally do not pertain (J&V 2008; Frijda 1993). Given the apparent lack of survival function in music-related emotion, the primary question J&V seek to answer is "how does music evoke emotions?" They claim the answer involves cognitive appraisal to some extent, as well as six key mechanisms that underlie emotional responses to music, briefly listed here to be explained in more detail in the next section: (a) brain stem reflexes; (b) evaluative conditioning; (c) emotional contagion; (d) visual imagery; (e) episodic memory; and, (f) musical expectancy.

J&V are correct in drawing attention to the lack of focus on underlying mechanisms of music-related emotions. However, they fail to notice that the six 'underlying' mechanisms they list are not the most 'basic' or 'fundamental' of mechanisms, and may be further grounded[2] in their relation to physiological response and homeostatic function,[3] which will be introduced later in this section, and further reviewed in part three.

What's Missing in the Search

Throughout the remainder of the paper, we will focus our attention on component (f) musical expectancy, and expand this notion to include physiological expectancy. In accord with Vuust & Frith (2008), I will claim that the combination of musical and physiological expectancy can be said to guide the first five of the proposed mechanisms. As will be elaborated throughout this paper, active interoceptive inference accounts of emotion, in conjunction with a neurofunctional model provided by the Quartet Theory of Emotion, provide a more nuanced explanatory account of both emotions in general and music-related emotional experience through an emphasis on the role of physiological

2. I take 'grounded,' in this sense, to refer to being further rooted in a more basic and fundamental mechanism.

3. Homeostasis refers to an organism's tendency to maintain its functioning with in a viable range, ensuring organism's survival. For example, the bodies of diabetic individuals fail to maintain homeostasis between levels of sugar and insulin without lifestyle intervention in the form of medicine or dietary changes.

homeostasis. Already, the six mechanisms provided by J&V can be linked through an underlying process of this sort, by cueing physiological and neural responses.

The Importance of Expectancy

In response to J&V, Vuust & Frith (2008) highlight the privileged role of expectancy in understanding music and its effects on emotion as an important component in a hierarchy of music-related emotional mechanisms. They claim that musical expectancy should be the most fundamental mechanism of the others in J&V's list. This is supported from a consensus among music theoreticians that musical experience and emotion is importantly conveyed by anticipation of local auditory events as well as of deeper musical and mental structures, such as overarching rhythmic and harmonic patterns, and contextualized by information about the piece and the memories of the listener. This view was originally explored in Meyer's 1956 work, "Emotion and meaning in music," and has been elaborated and extended to include physiological and neural processes of expectation in Huron's 2006 work "Sweet Anticipation." Accounts such as Meyer's and Huron's claim that expectation arises from culture dependent statistical learning, familiarity with a piece of music, short term memory for immediate musical history, and deliberate listening strategies (attention), and are consistent with the predictive processing framework. Vuust & Frith (2008) summarily claim that "the musical experience is dependent on the structures of the actual music, as well as on the expectations of the interpreting brain."

This response accurately points out that expectancy has a large and likely privileged role in the processing and perception of music-related emotions—however, it is incomplete to assume that so much emphasis should be placed on expectation related solely to the exteroceptive content in the auditory musical structure. Given the important role of physiological state in an emotional experience, expectancy theories of musical emotion should move beyond merely appealing to predictive processing of the structural features of the music itself. Instead, expectancies related to interoceptive/physiological states of the listener/performer themselves should be incorporated. As such, I will provide a more focused view of predictive processing of emotion in music below, emphasizing the role of active interoceptive inference. First, let's revisit our concert-goers:

Audience Spotlight 2

When we introduced Rylan and her grandfather, Henry, they were both attending the same concert and listening to the same music, however they had substantially different emotional responses. This means the music and context alone cannot serve as a one-to-one cue for a certain emotional state. Instead, maybe the music is cuing different extra-musical associations for our concert-goers which are themselves tied to different emotional responses. This remains an open possibility, however it does not help to fully explain the fundamental processes giving rise to their respective joy and unrest at the Rammstein concert. After a journey through some proposed underlying mechanisms, we paused on the important role of expectancy. Although Rylan and Henry are both listening to the same music in the same context and at the same time, they each have different musical expectations due to differing life and musical experiences. So, some of the difference in emotional response may be due to the mere exposure effect: Rylan has had more exposure to industrial metal music such as Rammstein and to the atmosphere of a Rammstein concert; thus, she has more precise expectations of the musical input and is more likely to enjoy it.

However, we are focusing heavily on exteroceptive information in justifying this difference in emotional experience, and we are analyzing the differences on a personal level. We still lack focus or explanation of the underlying neural and physiological mechanisms by which music effects emotional responses.

In the following sections, predictive processing accounts of emotion, including the active interoceptive account, as well as the neurofunctional mechanism proposed by the 'quartet theory' of emotion will be laid out, followed by their role in the processing of music-related emotions.

Predictive Processing Account of Emotion

Incorporating Interoceptive Expectancy

The idea that there is an interoceptive component to emotion is not new. Two-factor theories such as those proposed by James (1884), Lange (1885), and Schachter & Singer (1962) all discuss emotion as an appraisal of physiological arousal. However, these theories focus on feed-forward (bottom-up) models, where the sensory and autonomic systems accumulate evidence in the form of various physiological and arousal states, before being cognitively appraised and thus labeled as an emotion state. The

active interoceptive inference account of emotion, however, assumes that interoceptive processes are subject to the same reciprocal relation between top-down predictions and bottom-up stimuli as the exteroceptive and proprioceptive processes involved in perception and action under the predictive processing framework (Seth 2013). In predictive processing (PP), cascading predictions from top-down generative models are met with bottom-up prediction errors, which serve to either update the predictive models or to motivate action (Clark 2013).

PP models of exteroceptive signals enable inference regarding the states of affairs in the external world which are most likely to cause a set of sensory states through the process of prediction error minimization (PEM). PP models of interoceptive signals, however, serve control and regulation of physiological states with the goal of maintaining homeostasis (Seth 2013; Sel 2014; Seth & Friston 2016). Predictions in interoceptive inference, rather than generative models about an external state of affairs, are internal homeostatic 'set points,' toward which the activity of autonomic and affect systems is driven as a result of minimizing the difference (prediction error) between this set point and current physiological state (Seth & Friston 2016). This grounds interoceptive prediction, at the subpersonal level, in evolutionary goals toward maintaining homeostasis, which is essential for biological fitness, and fulfills the survival function of emotions emphasized by Frijda (1993). Importantly, active interoceptive inference alone does not account for every aspect of emotional content, but augments existing two-factor theories of emotion by contextualizing interoceptive predictions with concurrent active inference over proprioceptive and exteroceptive cues. This involves a 'cognitive' appraisal in terms of higher cognitive functioning (such as stimulus evaluation),[4] but this appraisal need not be conscious or linguistically achieved. In contrast, linguistic labeling of an emotion state involves a translation, reconfiguration, and ultimately an approximation of a much more specific and nuanced underlying process:[5] an underlying process which integrates multiple sources of predictive information. The components of a process of this

4. Stimulus Evaluation Checks (SEC) are outlined in the component process model of emotion (Scherer & Zentner 2001; Scherer 2009; 1999). SECs refer to a multi-level appraisal process consisting of sequences of appraisal checks of the appropriateness of a stimulus and associated emotion (akin to the level of prediction matching or prediction error) from lower to increasingly higher levels of perceptual and cognitive processes.

5. In addition to the lack of semantic precision of emotion terms to emotion events, another known reason for the ineffectiveness of studying emotion appraisal gained from linguistic self-report is that these reports "may contain appraisals that are part of the emotional response, rather than belonging to its causes" (Frijda 1993).

sort are outlined in the following table in regard to the predictive information available in music.

Music Specific Sources of Predictive Information	
Exteroceptive Inference: PEM over the causes of external physical stimuli • the music itself (acoustic stimuli) • observation of the performer(s) movement, appearance, etc • other's body movement (dancing, etc)	**Interoceptive Inference:** PEM over the causes of autonomic, physiological stimuli (cued by exteroceptive/proprioceptive information) • heartbeat, breathing rate, etc • skin conductance • arousal • hormone activity
The integration of which leads to music-related emotions.	

This non- (perhaps pre-) linguistic cognitive appraisal, in the form of stimulus evaluation checks and prediction error minimization, occurs over many levels, and is achieved when prediction error is minimized at the lowest possible level, contextualized by predictive information from a multitude of sources.[6] When the predictive model fits (what level of homeostasis should be expected) then the emotional response stabilizes (Gerrans 2017). Thus, emotional systems, such as those which will be elaborated in the Quartet Thoery of Emotion (Koelsch et al 2015), coordinate other perceptual and inferential systems in the task of determining self-relevance or survival value of a stimulus with information from both low-level affect, and high-level metacognition. Activity in the insular cortex 'allows us to *feel* how things in the world matter to us, in the form of affect' (Gerrans 2017), with primary interoceptive representations in the posterior insular

6. For the purposes of this paper we will largely focus on the exteroceptive information available in acoustic music stimuli to a listener. Proprioceptive information is gained from movement to music, as well as the physical act of producing music itself, and may enhance any of the expectation effects of exteroceptive information as well as provide more direct cues for interoceptive process (for example, playing the drums will do more to increase blood flow and breathing rate than mere listening because of the direct physical action involved.)

cortex, and secondary integrated experience in the anterior insular cortex (Seth 2013; Seth & Critchley 2013).

What's Missing

Applying active interoceptive inference to emotion does extend existing two-factor theories of emotion by incorporating an integration of prediction and prediction error minimization over exteroceptive, proprioceptive, and interoceptive stimuli; however, we are still left with a two-factor framework. Current physiological condition is merely replaced with current homeostatic condition of the body, followed by cognitive appraisal in at least the minimal sense of stimulus evaluation, and perhaps action to maintain homeostasis. Even so, this consideration of physiological homeostasis may provide us a bit more insight into the experience of our concert-goers, Rylan and Henry:

Audience Spotlight 3

The last time we saw Rylan and her grandfather, Henry, we guessed that some of the difference in their emotional responses may be due to their individual musical histories. Rylan and Henry have different musical expectations due to different musical experiences. Henry's range of possible musical expectations is relegated to traditional country tunes, while Rylan's range of possible musical expectations includes industrial metal music, so she has more precise expectations of the music of Rammstein. While this mere exposure effect can to some degree explain our concert-goers' relative familiarity and positive or negative affect with respect to the Rammstein concert, we still don't know what is going on at the subpersonal level. Why is the exteroceptive information apparently cuing different emotional responses? For this, we turn to interoceptive inference accounts of emotion.

In addition to expectations of the external musical stimuli, Rylan and Henry each have expectations relating to their own internal bodily states. So, some of the difference in emotional response is due to differing histories of physiological responses related to music listening. Rylan has had more experience with the effects of Rammstein's music and the atmosphere of their concerts on her internal states. Thus, she has more precise expectations of her body's response to their music, and is more likely to enjoy it.

The internal bodily expectations of Rylan and Henry are not necessarily cognitive expectations (such as, 'I expect loud music and fast rhythms to energize me'), but are rather subpersonal, interoceptive expectations of physiological homeostasis. Rylan

listens to heavy metal music quite frequently, so when the external music cues internal changes—rising heart rate, breathing rate, heightened levels of arousal etc.—this occurs relatively near the range of Rylan's homeostatic set point, but drastically out with the homeostatic expectations of Henry, causing him to feel agitated while Rylan instead feels energized.

We seem to have a decent explanation of the valence of similar arousal related emotions such as 'agitated' vs. 'energized.' However, incorporating active interoceptive inference still does not explain how the homeostatic range, or 'set point,' impacts nuanced emotions and *complex* social emotions, or how external musical stimuli can modify these physiological processes.

Active interoceptive inference accounts of emotion focus heavily on activity in the insular cortex. Within the insula is a viscerotopic map with general representations of interoceptive states, akin to the retinotopic map in your visual cortex. Relying on this viscerotopic map and ambiguously integrated experience within the insular and cingulate cortices—with offhand mention to other emotional systems—these interoceptive inference accounts lack a clear description of the active *mechanism* which is implemented in the maintenance of homeostasis occurring across multiple neural and bodily systems. I propose that this mechanism can be filled in to the active interoceptive inference accounts by incorporating a neurofunctional model of emotional experience, which more explicitly details how the brain and associated biological systems effect homeostatic maintenance and generates affective experience. In the following section, a plausible mechanism over which active interoceptive inference occurs will be explored in reference to the neurofunctional model put forth by the quartet theory of emotions. It is not yet immediately clear how much active interoceptive inference alone adds to theories of music-related emotions besides a more detailed description of emotions associated with arousal states. However, once we incorporate QTE as the active mechanism, we will revisit the differing emotional experience of our two concert-goers and the enhanced explanations of their exuberance versus unease, and differing feelings of belonging.

Part 3

Neurofunctional 'Quartet Theory of Emotion' and Predictive Processing

The 'Quartet Thoery of Emotion' is a neurobiological theory proposed by Koelsch et al. (2015) that links four classes of emotion to a quartet of neurobiological affect systems (brainstem-centered, diencephalon-centered, hippocampus-centered, and orbitofrontal-

centered), which interact in a dynamical way with biological effector systems (peripheral arousal; action tendencies; motor expression; memory and attention). Reciprocal interactions between these affect systems, effector systems, and conscious appraisal systems (as well as reciprocal interactions within elements of each system) provide a neurological mechanism by which an 'emotion percept' arises and becomes consciously attended to. I posit that these reciprocal interactions consist of predictive processes as *model-updating* features of the system (via the neurological affect systems) and *world-updating* features (via the biological effector systems). In the following section, I will outline each component of this neurofunctional model and demonstrate the role of active inference in each, extending active inference's role from merely interoceptive prediction error minimization in the insular cortex to prediction error minimization across a series of neural and biological structures—the integration of which, in varying combinations and degrees, results in emotional experience. Following this in-depth theoretical outline of the merging of PP and QTE accounts of emotion, I will show how the merging of these two approaches sheds insight into the subpersonal processes underlying the emotional experience of music in part four, while revisiting our concert-goers Rylan and Henry.

Model Updating: QTE Neurobiological Affect Systems

Brain-stem centered In order for active interoceptive inference to work toward the goal of improving physiological homeostasis in various visceral and autonomic systems, these systems must be somehow linked and integrated. At lower levels of processing, these physiological processes are linked and integrated already at the level of the brainstem, as well as the insular and anterior cingulate cortices indicated in predictive processing accounts of emotion. The brainstem is structurally and functionally implicated in relation to the auditory, vestibular, visceral, autonomic and parabrachial nerves. The brainstem and hypothalamus are also in the relevant structural location to receive and incite activation from/to neural pathways corresponding to interoceptive systems—including the insular cortex—and to integrate information from each system to form a cohesive emotional feeling state. These regions generate, modulate, and integrate somatomotor, visceromotor, and neuroendocrine activity essential for survival. As an important center for feelings of arousal, the brainstem regulates homeostatic activity throughout the body, including cardiovascular and hormone activity. Building on the brainstem's integral role in QTE, I posit that the brainstem is the key location for 'model-updating' activity in predictive processing of interoceptive states through tracking and adjusting the parameters of homeostatic 'set-points' given context from different bodily

and neural processes. The brainstem is also implicated in 'world-updating' through cuing the motor activity necessary to regulate these processes.

In music-related emotions, we can see brainstem-centered predictive processes at work at electronic dance music concerts, or raves. The DJ holds the crowd in ever building anticipation, adding one new level of instrumentation at a time, building tension through introducing increasing amounts of information (and building up prediction error), before ultimately reintroducing the bass beat and resolving the tension in the music and the prediction error in the brains and bodies of every member of the audience. At each level of increased musical tension, the arousal and anticipation felt by the audience is increased because the amount and rate of deviation from homeostatic arousal norms is increased concurrently with the increasing amount and rate of musical change. This homeostatic arousal state is reset, or updated, as soon as the beat drops and the music and body are brought back in line with the listener's expectations.

Diencephalon centered Much of the work on active interoceptive inference highlights the role of the anterior cingulate cortex and the anterior insular cortex in the integration of prediction and prediction error from specifically interoceptive processes. However, the brainstem, as expressed above, as well as the diencephalon centered systems both integrate information from the body and other brain systems themselves. The diencephalon centered system is associated with the dopaminergic reward system and contains the thalamus, which is associated with perceptual aspects of pain, and the hypothalamus, which is associated with behavioral, autonomic, and endocrine activity and perceptual aspects of pleasure and fun. Information from all of the senses passes through the thalamus, and, given contextual information from the orbitofrontal cortex, becomes associated with affective valence before conscious perception. The hypothalamus processes: (1) homeostatic needs and fulfillment, incentive stimuli (potential to fulfill needs); threatening stimuli, novel stimuli; as well as, (2) input from other affect systems (such as the brainstem, hippocampus, OFC as well as the amygdala, anterior cingulate cortex and anterior insular cortex). These processes within the hypothalamus are important for ensuring the appropriateness of an emotion given the external, environmental context as well as internal bodily context.

I posit that the appropriateness of an emotion in the QTE—or at least of a certain physiological state—is determined by both the amount of prediction error (the total amount of deviation from homeostatic norms), as well as the expected rate of prediction error (the rate at which this deviation occurs, as well as resets). At our rave, the peak emotion state and ultimate release of tension at the drop, or reintroduction of the beat, not only resets homeostatic arousal states but also results in

activation of the dopaminergic reward system within the diencephalon-centered affect system as the world and the model, the music and internal bodily state, now match the listener's expectations.

Hippocampus centered Thus far, we have examined how active interoceptive inference occurs at lower-level perceptual processing in maintaining physiological homeostasis. Informed by interaction from the brainstem and diencephalon centered systems, active inference is also occurring at 'higher' cognitive levels of processing. These higher levels of processing extend toward processes such as memory and social behavior, yet are rooted in interoceptive homeostatic aims associated with satiating emotions, or evolutionarily beneficial non-satiating emotions such as attachment-related emotions—both associated with hippocampal activity which will be discussed in more detail shortly.

The hippocampus has dense reciprocal connections with other structures that regulate behaviors essential to survival. These dense reciprocal connections enhance the hippocampus' structural relation to a complex network of emotion systems, and lend structural support to the interaction between memory and emotion. Hippocampal activity is less directly associated with fulfilling immediate homeostatic needs and is more associated with long term attachment related affects, which are implicated in social interaction, sense of belonging, and social cohesion.

I posit that *repeated* social interactions, such as group music making or listening, which *individually* enhance the maintenance of homeostatic needs (as well as positive reward and arousal from the dopaminergic system and brainstem respectively) builds up a predictive model associating higher-level social activity with cascading prediction error minimization down through lower-level physiological homeostatic maintenance. The attachment-related affects which result from this socially cued cascade of PEM results in the positive feelings of belonging. Consider that the rave we've been discussing is actually one of a weekly series of concerts. A group of people have been attending and enjoying these raves every Saturday night for some number of months. Although they don't otherwise know each other, these repeated social interactions, which have individually resulted in various low-level positive affect, have now come to develop in each of the rave-goers a deep sense of attachment and belonging.

Orbitofrontal cortex centered This system most clearly corresponds to the concept of cognitive appraisal; however, the OFC is not a language area. The OFC is responsible for forming concepts and norms, which are 'propositionally not available' or unconscious. The OFC evaluates external and internal stimuli, as well as information from other affect systems for reward/punishment potential and response 'by indicating vegetative,

neuroendocrine, behavioral and cognitive programs according to social requirements and social norms' which are learned early in life (Koelsch et al 2015).

Predictive models of physiological homeostasis at lower levels are both contextualized by and contextualize higher level cognitive systems centered on the hippocampus and orbitofrontal cortex. For example, long-term attachment related affects generated by the hippocampus will be associated with predicted levels (or a certain homeostatic range) of dopaminergic and arousal levels in the diencephalon and brainstem centered systems, respectively, which are themselves associated with a specific homeostatic range of interoceptive bodily states. The predictive models maintained by the hippocampus and the OFC will be models developed early in life, and continually developed throughout one's lifespan.

In the arena of music-related emotions, I posit that 'lower'-level predictive models more directly associated with physiological homeostasis (such as the brainstem-centered and diencephalon centered systems) will be more immediately impacted by current external musical cues. However, the habitual expectations associated with these 'higher'-level models will be less susceptible to the immediate effects of musical cues and will be most impacted due to repeated experience with music listening and group music making. Consider our rave-goers on one Saturday when the music of the particular DJ was lackluster and unfulfilling. Although the music that night did not result in the same positive low-level affect, in the form of either homeostatic maintenance or reward activation, the rave goers still maintained a sense of attachment and belonging developed by months of rave attendance. In addition, the rave goers shared a sense of disappointment relating to their music listening experience that they could not quite articulate, as a result of the deviation from the OFC's higher level (non-propositional) concept of a 'good' rave vs. the lackluster event on this night.

World Updating: QTE Neurobiological Effector Systems

The affect systems above concerned localized neurological systems which mediate, interpret, or control other bodily processes and generally constitute 'model-updating' processes. The effector systems are those bodily processes which can perform action in the world; they fulfill the 'active,' 'world-updating' component of active inference, and can act to bring the body and world in line with homeostatic expectations generated by the activity of the neurological affect systems. The four emotional effector systems are: motor systems, peripheral physiological arousal systems, attention systems, and memory systems. Information and action from all four of these systems contextualizes

information and action from the other effector systems, and are variously integrated within the neurological affect systems described above. Motor systems govern action tendencies (skeletal and muscular activity related to behavior) as well as the expression of emotion (through facial expression or vocalization). Peripheral physiological arousal systems modulate endocrine activity, vegetative systems (changes in sympathetic/parasympathetic activity), as well as motor and non-motor activity of all organ systems (motor activity including heart activity, breathing, vasoconstriction/dilation; non-motor including immune function, wound healing, energy metabolism). Attention systems can include motor activity such as head turning and eye gaze, as well as non-motor activity concerned with cognitive/psychological attention. And finally, memory systems monitor the selection of information for long and short-term storage, as well as access to that information.

Each of these effector systems perform the actions necessary to minimize prediction error within the neurobiological affect systems. For instance, in our rave example, attention and motor systems are engaged as the listeners attentively wait on input from the DJ, increasing their movement in conjunction with the increased activity of the music until finally jumping in sync with the reintroduction of the bass. The movement of their body then corresponds with the movement of the music, actively minimizing prediction error between their body and the environment. In part four, we will investigate this phenomenon in more detail, as well as how our interlude to the world of electronic music have actually helped us to explain the emotional experience of our heavy metal concert-goers.

Part 4

Enhanced Explanation of Emotional Experience in Music

Audience Spotlight 4

Since we've met Rylan and her grandfather, Henry, we have attributed some of the difference in their emotional responses to the Rammstein concert to differences in exteroceptive expectations related to their individual musical histories, as well as differences in interoceptive expectations related to their experiences of music-associated physiological responses. Henry feels *agitated* both because of his unfamiliarity with Rammstein's music, as well the unfamiliarity with how their music makes him feel. In contrast, Rylan feels *energized* because she has expectations that encompass both Rammstein's music, and how she feels when listening to their music.

Their more nuanced emotional experience of unease or exuberance is impacted by the amount and rate at which any unfamiliarity or expectation changes *over time*. As Rylan's expectations are continually being met for both the musical and internal stimuli, her energized feeling turns to *exuberance*, especially at peak musical moments. As Henry's deviation from physiological homeostasis is continually increasing, his agitation morphs to unease.

As a bonus, Rylan's repeated Rammstein concert attendance reinforces these energized, exuberant feelings and contributes to a sense of belonging within the Rammstein crowd. Henry's repeated Merle Haggard concert attendance has the same effect in the context of country music fans, but he has no such built up, positive, contextual associations for Rammstein concerts. Thus, Rylan is more likely to feel *a sense of belonging* when listening to Rammstein's music and attending their concerts.

I have established that integrating the neurofunctional Quartet Theory of Emotion with the predictive processing accounts of emotion outlined by active interoceptive inference gives us a more clear and nuanced picture of the subpersonal activities underlying emotional experience than any one account alone. Below, we will consider the enhanced explanations provided by integrating these accounts in the case of emotional experience associated with: (a) basic arousal; (b) nuanced emotional states; and, (c) complex social emotions, expanding on the short vignettes of Rylan and Henry.

Basic Arousal

Active interoceptive accounts of emotion rely heavily on the viscerotopic map within the insular cortex, but as I have demonstrated in part three, more attention must be paid to other neural systems implicated in interoceptive and emotional processing. Music, as an exteroceptive cue for interoceptive activity, already affects physiological arousal at the level of the brainstem (Koelsch 2014). The location of the reticular formation, a part of the brainstem which is implicated in maintaining arousal and homeostatic functioning, is at a structurally advantageous location to mediate and integrate information from the cochlear and vestibular nerves. Music, as an auditory event, activates both of these nerves through acoustic signals and affects the movement of fluid within the vestibular system. In fact, this vestibular fluid movement partially explains a drive to move your head along to the beat of a particularly rhythm of bass heavy song (Phillips-Silver 2009). I posit that this movement along to a beat corresponds to the world-changing, active inference of matching physiological vestibular state by activating motor effector systems to move your body or head to the external, musical stimulus.

If you will recall our rave, listeners moved in sync to whatever beat the DJ provided, ensuring that the movement of their body matched the activity in the world (the beat of the music). As Rylan headbangs along with the beat of *Du Hast*, she is actively responding to the energy of the music and (unknowingly) to the movement of vestibular fluid in her ears via coordinated activity between the brainstem-centered affect system and motor effector systems. This minimizes prediction error through matching the movement of fluid in her ears with the expected correlated movement of her body, maintaining an energized emotional state. Henry, however, standing still in staunch rebellion is remaining in a state of unresolved prediction error, maintaining his agitation.

In addition to communication with the vestibular nerve, the reticular formation in the brainstem is in a structurally advantageous location to transmit this auditory phenomenon of music toward regions of visceral and autonomic processing, including the insular cortex, contributing to the experience of goosebumps, or frisson, in peak emotional experience of music.

Nuanced Emotions

While the insular cortex of PP accounts of emotion tracks and integrates specifically interoceptive processes, the brainstem-centered and diencephalon-centered systems of

QTE perform integration of input from a variety of other affect systems.[7] The reward circuit within the diencephalon-centered system is particularly activated in response to peak emotion in music, and is associated with a goose-bumps sensation called 'chills' or 'frisson' (Salimpoor et al. 2011). As part of the peripheral physiological arousal effector system, frisson may occur as a result of the fulfillment, or release, of built up tension within a musical piece. I posit that this occurs because the world (the music) finally matches the expectation of the listener, and the sudden minimization of exteroceptive prediction error, results in the physiological response of goosebumps.

In the middle of *Du Hast*, all of the back-up instrumentation including the heavy beat of the guitar and drums disappears, leaving only the lead singer, a light synthesizer, and a sudden wealth of prediction error for a period of about fifteen seconds. When the beat 'drops' (to use the terminology of our rave example), Rylan and the other Rammstein fans begin again to jump and headbang along to the beat. Rylan experiences chills in reaction to this sudden minimization of exteroceptive prediction error and resetting of homeostatic arousal states at the peak emotional moment of the song. Henry does not have this reaction, perhaps because, when the background instrumentation dropped out, he expected the song to be ending. In this case, the sudden reintroduction of the heavy beat instruments gives Henry an unexpected, and unappreciated surprise, increasing the deviation from his physiological homeostasis.

Henry's agitation, over time, is enhanced through not only an increased amount of prediction error but also the high rate of prediction error as his body continues to deviate from its levels of homeostatic norms, even after peak moments in the music. This persistent prediction error, over time, not only causes him to feel a general negatively valenced arousal state but also contributes to Henry's nuanced emotional experience of unease. In contrast, Rylan's persistent minimization of prediction error, as well as return to an expected homeostatic range after peak moments, causes her to feel not just a positively valenced arousal state, but contributes to her nuanced emotional experience of exuberance.

Complex Social Emotions

Music may affect immediate survival functions through homeostatic autonomic activity and activation of the dopaminergic system (the reward circuit), as outlined in part three concerning the brainstem and diencephalon-centered systems, but it may also

7. Including the brainstem, hippocampus, orbitofrontal cortex, as well as the amygdala, the anterior cingulate cortex, and the anterior insular cortex (Koelsch et al. 2015).

contribute to long term survival functions by enhancing the effect of attachment related emotions through group music making or group music listening.

The habitual predictions of the hippocampus-centered and OFC-centered systems contributes to higher level habitual predictions which are learned early in life. These habitual predictions are associated with the socio-cultural norms to which individuals are exposed early in life, resulting in firmly-established models of such phenomena as the language and music they hear, and social environments in which they interact. If higher-level expectations are met *at the same time* as the lower-level physiological homeostatic expectations are met, then prediction error among these quartets of systems are being maximally minimized, leading to a sense of group belonging.

Rylan has been listening to music like Rammstein since she was in grade school, and she has well-developed habitual predictions surrounding heavy metal music, metal concerts, and metal fans. In addition, the positive effect of their music on her physiological homeostasis and corresponding emotional experiences has been well established. Because prediction error is being minimized at both higher-level conceptual processing, as well as lower-level physiological processing, at this Rammstein concert, Rylan feels as if she is *right where she belongs*.

Henry, however, has not grown up in an environment surrounded by metal music and has developed none of the firmly established models of the social environment surrounding metal music. Neither his higher-level conceptual expectations, nor his lower-level physiological expectations are being met, causing Henry to feel *as if he distinctly doesn't belong* at this Rammstein concert and around these Rammstein fans. Even if it is the case that Henry logically expects a certain social environment at this concert, and has developed a higher-level conceptual model of a metal concert environment, because this higher-level prediction error is not being met *at the same time* as lower-level homeostatic prediction error, he does not have all of the necessary ingredients for a feeling of belonging.

Part 5

Conclusion

Throughout this paper, I have demonstrated that the Quartet Theory of Emotion fills many of the holes left in predictive processing framework, specifically regarding the PP account of emotion. The QTE provides a neurofunctional model detailing which neural and bodily systems are carrying out active inference over which combination of neural

and physiological homeostatic processes associated with varying types of emotional experience. This provides a detailed mechanism for active interoceptive inference to contribute to the formation of basic arousal states, nuanced emotional states, and complex social emotions.

Objections

The conscientious reader might draw the objection that although QTE fills in explanatory gaps within predictive processing, what does the PP account of emotion add to QTE? Might the quartet theory itself be a sufficient explanation of emotional experience without appealing to PP? This is ultimately an empirical question with ramifications for the overarching framework of predictive processing itself. If it is found that QTE covers all aspects of emotional experience, with the involvement of perhaps some other predictive process, then it will contribute to identifying at least one boundary on the limits of explanatory power of predictive processing.

However, if QTE does implement predictive processing in each of its sub-systems, then this provides further evidence toward PP as an overarching framework for the embodied brain. Indeed, the QTE as it stands already seems to incorporate the model- and world-updating features inherent in the active inference aspect of PP via the activity of the affect and effector systems respectively. These systems in QTE maintain lower-level physiological homeostasis as well as higher-level conceptual expectations involved in forming the varying levels of emotional experience—which is exactly what a full-fledged predictive processing account of emotions aims to describe. It is not only parsimonious to integrate PP and QTE approaches to explain emotional experience but also appealing to the PP framework provide a seamless integration of the personal-level phenomenological experience with the sub-personal level of the underlying neural and bodily processes, as demonstrated through the running vignette of Rylan and Henry and the personal and sub-personal differences in their emotional experiences while attending a Rammstein concert.

In this paper, I have discussed how the predictive processing framework expands upon traditional, bottom-up, two-factor theories of emotion as passive physiological evidence-building and subsequent cognitive appraisal (James, 1884; Lange, 1885; Schachter & Singer 1962) by incorporating active inference based on predictive models of the causes of external *and* internal stimuli (Seth & Critchley 2013; Seth & Friston 2016). Thus, emotional content from music is evoked as a result of active exteroceptive inference related to the physical musical stimuli (based on statistical regularities of the current

musical event and past experience) as well as active interoceptive inference regarding the listener's current autonomic, and physiological states. I have demonstrated that this is a productive route toward explaining the differing feelings of our concert-goers, Rylan and Henry. In addition, I have proposed that this general predictive processing framework is implemented through a 'quartet' of neurofunctional mechanisms (Koelsch et al 2015), which are dynamically implicated in the neural and physiological processes underlying general and music-related emotional experience.

Conceptualising emotion as active inference over both external and internal processes, implemented and maintained through a dynamically interacting subset of neural pathways offered by QTE, provides both a mechanism by which music evokes emotion and results in the subjective difference in the reported emotional experience of music between individuals, based on varying levels of neural interactions which arise from different prior experiences informing both their exteroceptive and interoceptive predictive models.

Avenues for Future Research

A more complete theory of the neurofunctional and physiological processes involved in emotional experience, such as that offered by the integration of the quartet theory of emotion and predictive processing, can help to solidify the boundaries, or demonstrate the continuity of emotional and cognitive processes. In addition, by expanding the QTE+PP model to incorporate distinctly social activities such as group music making, we can enhance our understanding of how social interactions affect both emotion and cognition in an increasingly second-person neuroscience. That is, a neuroscience that takes into account the inherently social and interactionist nature of the mental and behavioral activity of its participants, rather than merely testing how participants *simulate* or *theorize* regarding the intentions and activity of other minds.[8] Emphasis on interoceptive processes, and the feelings of belonging in group musical experience can further develop

8. A second-person neuroscience assumes greater influence of the Interaction Theory of Mind, understanding that others are experienced as a subject through mutual emotional and embodied engagement. In contrast, current first-person and third-person approaches to neuroscience assume that individuals take other minds as objects, either simulating the intentions and actions of other minds through a mentalizing or mirror neuron network (Simulation Theory of Mind), or undergoing a complex series of inferences (Theory-Theory of Mind). The experimental paradigms attached to these first- and third-person accounts leave no room to demonstrate the importance of interaction in social cognition (see Schilbach et al. 2013 for further discussion).

research into the concept of the self and the extended self, and whether selfhood lies within our body or brain or emerges through interaction in our social environment.

References

Clark, Andy. 2013. "Whatever next? Predictive Brains, Situated Agents, and the Future of Cognitive Science." *Behavioral and Brain Sciences* 36: 181-253.

Clark, A. 2016. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford, United Kingdom: Oxford University Press.

Huron, D. 2006. *Sweet Anticipation*. Cambridge, MA: MIT Press.

Frijda, N. H. 1993. "The place of Appraisal in Emotion." *Cognition and Emotion* 7 (3/4): 357–387.

Gerrans, P. (Presenter). 2017, June 29. *Emotions and Subjective Perspective*. Speech presented at The Human Mind Conference, The Møller Center, Cambridge, United Kingdom.

Hohwy, J. 2013. *The Predictive Mind*. Oxford, United Kingdom: Oxford University Press.

James, W. 1884. "What Is an Emotion?" *Mind 9*: 188-205.

Juslin, P. N., & Västfjäll,D. 2008. "Emotional Responses to Music: The Need to Consider Underlying Mechanisms." *Behavioral and Brain Sciences* 31: 559–575.

Koelsch, S. 2013. "Emotion." In *Brain and Music*, 203-240. Hoboken, NJ: Wiley-Blackwell.

Koelsch, S. 2014. "Brain Correlates of Music-Evoked Emotion." *Nature Reviews Neuroscience* 15: 170–180.

Koelsch, S. et al. 2015. "The Quartet Theory of Human Emotions: An Integrative and Neurofunctional Model." *Physics of Life Review* 13: 1–27.

Lange, C. G. 1885. "The Mechanism of the Emotions." In *The Emotions*, edited by D. Dunlap, 33–92. Baltimore, MD: Williams & Wilkins.

Meyer, L. B. 1956. *Emotion and Meaning in Music*. Chicago, IL: The University of Chicago Press.

Narmour, E. 1990. *The Analysis and Cognition of Basic Melodic Structures: The Implication Realization Model*. Chicago, IL: University of Chicago Press.

Phillips-Silver, J. 2009. "On the Meaning of Movement in Music, Development, and the Brain." *Contemporary Music Review* 28 (3): 293–314.

Salimpoor, V. N., Benovoy, M., Larcher, K., Dagher, A., & Zatorre, R. J. 2011. "Anatomically Distinct Dopamine Release During Anticipation and Experience of Peak Emotion in Music. *Nature Neuroscience* 14: 257–262.

Schachter, S., & Singer, J. E. 1962. "Cognitive, Social, and Physiological Determinants of Emotional State." *Psychological Review* 69: 379–399.

Scherer, K. R. 1999. "On the Sequential Nature of Appraisal Processes: Indirect Evidence from a Recognition Task." *Cognition and Emotion* 13: 763-793.

Scherer, K. R. 2009. "The Dynamic Architecture of Emotion: Evidence for the Component Process Model." *Cognition and Emotion* 23 (7): 1307–1351.

Scherer, K. R., & Zentner, M. R. 2001. "Emotional Effects of Music: Production Rules." In *Music and Emotion: Theory and Research*, edited by P. N. Juslin & J. A. Sloboda, 361–392. New York: Oxford University Press.

Schilback, L., et al. 2013. "Toward a Second-Person Neuroscience." *Behavioral and Brain Sciences* 36: 393–462.

Sel, A. 2014. Predictive Codes of Interoception, Emotion, and the Self [Review *Interoceptive inference, emotion, and the embodied self*, by A. K. Seth]. *Frontiers in Psychology* 5: 1-2.

Seth, A. K. 2013. "Interoceptive Inference, Emotion, and the Embodied Self." *Trends in Cognitive Sciences* 17 (11): 565–573.

Seth, A. K., & Critchley, H. D. 2013. Extending predictive processing to the body: Emotion as interoceptive inference [Review *Whatever next? Predictive Brains, Situated Agents, and the Future of Cognitive Science*]. *Behavioral and Brain Sciences* 36: 47–48.

Seth, A. K., & Friston, K. J. 2016. "Active Interoceptive Inference and the Emotional Brain." *Philosophical Transactions B* 371 (1708).

Van de Cruys, S. 2017. Affective Value in the Predictive Mind. In *Philosophy and Predictive Processing*, edited by T. Metzinger & W. Weise. Frankfut am Main: MIND Group.

Vuust, P., & Frith, C. D. 2008. "Anticipation is the Key to Understanding Music and the Effects of Music on Emotion [Review *Emotional Responses to Music: The Need to Consider Underlying Mechanisms*, by P. N. Juslin & D. Vastfjall]." *Behavioral and Brain Sciences* 31: 599–600.

cognethic.org