Productive Theory-Ladenness in fMRI

M. Emrah Aktunc
Ozyegin University

> We must suppose a very delicate adjustment whereby the circulation [of blood] follows the needs of cerebral activity. Blood very likely may rush to each region of the cortex according as it is most active, but of this we know nothing. I need hardly say that the activity of the nervous matter is the primary phenomenon, and the afflux of blood its secondary consequence.
>
> William James, 1890 (p.99)

1. Introduction:

As the above quote from James's classic *The Principles of Psychology* shows, the existence of a relationship between cognitive activity and patterns of cerebral blood flow is not a new postulate. However, a century of research in diverse fields had to be conducted before the techniques that made it possible to study this relationship became available. Several developments to achieve various scientific and technical goals took place over a century and eventually gave us functional magnetic resonance imaging (fMRI).[1] Techniques such as fMRI, as well as others (e.g. positron emission tomography (PET)), gave rise to cognitive neuroscience as a new research program. However, since its beginnings, this program has not been without its skeptics. Here, I argue that the sound use of fMRI in cognitive neuroscience gives rise to a productive kind of theory-ladenness thanks to its solid foundations in the physics of magnetic resonance and the physiology of hemodynamics. As I will demonstrate with concrete examples, this productive theory-ladenness makes possible the identification and control of potential errors in fMRI experiments (e.g. the development and use of the NPAIRS framework) and gives rise to novel and data-driven models of the neural bases of cognition independently of large-scale theories of cognition (e.g. the HERA model of memory). Before moving on, a brief summary of the skeptical accounts of neuroimaging will be useful.

[1] I focus on fMRI because it is the most commonly used neuroimaging method, but parallel analyses can be made about other neuroimaging methods, e.g. MEG or fNIRS, *mutatis mutandis*.

Most skeptical accounts of neuroimaging conclude that hypotheses of cognitive neuroscience are underdetermined by neuroimaging data. This body of work can be broadly classified under two groups: one group raises doubts about the theoretical significance of neuroimaging results; and the other points to methodological difficulties of obtaining reliable inferences from neuroimaging data.[2] The work in the former group mostly appeals to issues of theory-ladenness and the work in the latter group emphasizes two major aspects of neuroimaging methodology: one, the highly complex nature of neuroimaging experiments; and two, the difficulty of modeling and drawing reliable inferences from neuroimaging data.

In his early critical account, Uttal (2001, 2002a, 2002b) cites ladenness of neuroimaging findings in modularist theories as the most problematic aspect of cognitive neuroscience. Uttal's main issue is whether or not we can accurately define isolated modules of cognition, which can then be localized in the brain. Regardless of the localizability of cognitive processes in the brain, their inaccessibility has long been a major problem in psychological science. Uttal cites as support for his arguments a well-known discussion of this problem by MacCorquodale & Meehl (1948); many cognitive processes assumed to separately exist are in fact hypothetical constructs, which may not exist at all as described by the researcher. These criticisms may not be totally valid for contemporary neuroimaging work employing newer approaches of interpretation and analysis, e.g. multi-voxel pattern analysis, neural fields, pattern-decoding reverse inference (for a recent review and discussion please see Nathan & Del Pinal, 2017).

Although Bechtel (2002a, 2002b) agrees that serious inferential problems exist in neuroimaging, he rejects Uttal’s general conclusion. He reminds that cognitive neuroscience is only one of the disciplines that proposes and tests theories that decompose a complex system into its components. In this kind of program, any and all such proposals are hypothetical; many of them are found to be incorrect and consequently revised. The important question, according to Bechtel, is not about the truth or falsity of modularist versus non-modularist theories, but whether or not researchers can obtain new findings from neuroimaging that would support certain componential theories while refuting or necessitating revision in others. In contrast to Bechtel, Hardcastle and Stewart (2002) agree with Uttal's skepticism and state that cognitive neuroscientists' modularist assumptions are "radically false" and neuroimaging cannot provide any support for modularist theories. This is because, prior to collecting data, researchers already assume the existence of

[2] The literature on the epistemology of neuroimaging includes a considerable number of works by philosophers of science and cognitive scientists. Naturally, a comprehensive review of this literature is beyond the scope of this paper.

specific cognitive processes localized in different parts of the brain, which is no more than a mere prejudice (ibid.). Thus, there is an inherent circularity in neuroimaging, which presumably destroys the potential of findings to support modularist theories.

Cognitive scientists have also written on inferential difficulties in neuroimaging. However, similarly to the philosophers above, they have focused on questions about the validity of inferences from data to theories of cognition. The debate between Henson (2005) and Coltheart (2006, 2010) provides an illustrative example; in essence, Henson believes that neuroimaging results can be used to adjudicate between competing theories of cognition and Coltheart directly rejects this.[3] Roskies (2009, 2010b) has provided strong defenses against Coltheart and other critics of neuroimaging.

The major theme in the above discussions is that the theory-ladenness of findings in neuroimaging creates serious problems of circularity in cognitive neuroscience. This theory-ladenness occurs when researchers interpret neuroimaging findings in terms of theories of cognition, which come with certain ontologies and assumptions about cognitive architectures and processes. However, these are not the only theories involved in fMRI research, which can be clearly seen when one looks at the scientific foundations of fMRI. The development of fMRI includes two separate trajectories of research; one in physics and the other in physiology. Research in these trajectories had progressed unconnected for decades from the 1920s to the 1990s and eventually converged to give rise to fMRI. The trajectory in physics had led to the discovery of the phenomenon of nuclear magnetic resonance (NMR) and its later application in the development of magnetic resonance imaging (MRI). The one in physiology had led to the discovery that changes in cerebral blood flow could be measured using MRI as the blood-oxygen-level-dependent (BOLD) response. These bodies of work had produced well-established and reliable scientific and technical knowledge which served as the foundations of fMRI. In addition, the techniques necessary for processing, filtering, and modeling raw data were provided by mathematics (e.g. Fourier transforms), advanced programming, and statistics.[4]

I argue that these theoretical and technical foundations give rise to a different kind of theory-ladenness in fMRI, which is productive because it enables fMRI studies to generate reliable data on the relationships between hemodynamic processes and cognitive task performance.

[3] Colin Klein (2010) provides a concise review of similar works by cognitive scientists who disagree on the epistemic value of neuroimaging.

[4] For a detailed account of the history of the development of fMRI see, for example, Aktunc, 2011, Le Bihan, 2014.

2. The Two Senses of Productive Theory-Ladenness in fMRI:

Productive theory-ladenness in fMRI enables researchers to:

> (a) identify, and control for, different types of methodological and inferential errors that may arise in the use of fMRI, and,
>
> (b) represent and investigate neural and cognitive phenomena in terms of hemodynamic data generated by fMRI as subjects perform well-defined tasks.[5]

The above statements can be unpacked as follows:

In an fMRI experiment:

> (1) Magnetic resonance phenomena, P_m , and hemodynamic phenomena, P_h , are used to investigate relationships between neural phenomena, P_n , and cognitive phenomena, P_c.
>
> > (1.1) Connections between P_m, P_h, and P_n have been previously well-established by evidence from physics of magnetic resonance and physiology of hemodynamics.
> >
> > (1.2) Connections between P_n and P_c have been previously established by neurobiological and neuropsychological evidence.
>
> (2) Theories of magnetic resonance, T_m , and theories of hemodynamics, T_h , are utilized to build the fMRI scanner that provides data on hemodynamic phenomena related to cognitive task performance.[6]
>
> (3) The terms and techniques derived from T_m and T_h plus computational and statistical techniques, are used to identify, formulate, and control for errors that may potentially jeopardize inferences in an fMRI experiment.
>
> (4) Thanks to 1 - 3, researchers can obtain reliable experimental data on relationships between P_n and P_c.
>
> (5) Thanks to 1 - 4, relationships between P_n and P_c can be represented and investigated using the terms of T_m and T_h in methodologically robust experiments.

[5] The term 'productive theory-ladenness' is not necessarily novel; for example, Chang (2012) uses it in his discussion of Hanson. However, in that context, theories are productive in the sense that they give "intelligibility to observations" by providing a conceptual framework and auxiliary assumptions (p.89). I use the term in a different, more methodological sense in which the knowledge and control of potential errors are productive in establishing the reliability of findings. It is also important to note that what makes this kind of theory-ladenness productive is not interdisciplinarity per se but precisely (a) and (b) above.

[6] Of course, in an fMRI experiment, theories of magnetic resonance and hemodynamics are complemented with computational and statistical techniques (e.g. Fourier transforms, spatial and temporal filtering, multivoxel pattern analysis, etc.) to preprocess, model and analyze data, which also constitute part of the productive theory-ladenness described here.

Propositions (2) and (3) above refer to the foundations of fMRI coming from theories of magnetic resonance and theories of hemodynamics, which establish that fMRI findings are laden in these theories and that productive theory-ladenness is a kind of theory-ladenness. Conclusions (4) and (5) above establish this kind of theory-ladenness as a productive kind because it is generative of any robust fMRI data set before issues of inference to cognitive functioning enter the picture. Regardless of what theoretical/computational approach a cognitive scientist employs or what type of inference a researcher decides to employ on a data set (e.g. adopting different types and/or techniques of reverse inference, see Nathan & Del Pinal (2017)), the first necessity is the generation of a reliable data set free of methodological flaws. This is made possible thanks to productive theory-ladenness. Thus, all who rely on fMRI data for their research must also rely on this productive theory-ladenness, which admittedly is quite different from the classic sense of theory-ladenness often cited to support skepticism of neuroimaging. So, it is necessary to delineate this concept of productive theory-ladenness in the context of previous discussions on theory-ladenness in philosophy of science.

3. Duhem, Kuhn, Hanson and Productive Theory-Ladenness:

Claims to the effect that experimental data are theory-laden abound in philosophies of science since the widely influential works of Kuhn (1962/2012) and Hanson (1958), who famously stated that "seeing is a 'theory-laden' undertaking" (p. 19), meaning that our observations of objects are laden with our prior knowledge of those objects. Although Kuhn and Hanson discuss some examples of research that feature observations generated by scientific equipment, such as microscopes or telescopes, their arguments are mainly about observation as a perceptual process. Yet, most contemporary scientific observation is done through the use of complicated instruments, such as electron microscopes, DNA sequencers, or fMRI, which work on the well-established principles provided by experiments and theories. As such, one could say that contemporary scientific observation is more theory-laden compared to earlier times.

Admittedly, the term theory-ladenness is ambiguous but Heidelberger (2003) usefully disambiguates and defines three different kinds of theory-ladenness, each of which he associates with Hanson, Duhem, or Kuhn. Heidelberger argues that, according to Hanson, apart from ladenness in human perception, theory-ladenness in science occurs when a theory is used to establish causal connections between observed phenomena. In this sense, theory-ladenness for Hanson is causality-ladenness: "The notions behind 'the cause x' and 'the effect y' are intelligible

only against a pattern of theory, namely one which puts guarantees on inferences from x to y." (Hanson, 1958; p. 64) For example, when people have difficulty recalling lists consisting of similar sounding words, namely the phonological similarity effect, this can be explained by a theory which posits that words are encoded and represented in memory in an acoustic or phonological format, which interferes with the recall of similar words (e.g. Baddeley & Hitch, 1974; Baddeley, 2003). The effect of impaired recall of similar sounding words would be causally explained by the theoretical posit that these words are acoustically or phonologically encoded.

Heidelberger (2003) then distinguishes Duhem's conception of theory-ladenness from Hanson's. For Duhem, theory-ladenness is not about establishing causal connections, but it occurs when the terms and relations proposed by a theory are used to state and interpret observed phenomena in an abstract and symbolic structure. However, not every field of science may be advanced enough to have produced theories with such symbolic or mathematical structures; so Heidelberger adds, when Duhem talks about theory-ladenness he seems to have in mind physics and its theories and not, for instance, physiology.[7]

Following Hanson and Duhem, Heidelberger (ibid.) continues with Kuhn. Theory-ladenness, for Kuhn, is "paradigm-ladenness" where a researcher can make sense of experimental phenomena only in terms of the theory (or theories) in the normal-scientific tradition in which she works. To elaborate, Heidelberger asks, when do anomalies, i.e. experimental findings unexpected by the paradigm, occur? What makes them anomalies? Is it because they do not fit a theoretical and abstract structure (*a la* Duhem) or they require the proposition of new causal connections (*a la* Hanson)? In other words, where do paradigm-induced expectations regarding experimental results come from? Heidelberger argues that if we look at examples discussed by Kuhn, we see that it is theoretical interpretation, or assimilation to theory, rather than causal understanding, that takes precedence in determining whether or not a series of findings are deemed anomalous for a given paradigm. On this point, Heidelberger distinguishes yet another possible meaning of theory-ladenness, which he calls "theory-guidance." Put specifically, this relates to "how the disposition to make a particular observation depends on the theoretical background [of a researcher]" (p.144). He argues that because this disposition is irrelevant to the "meaning of observation sentences" it must not be understood as genuine theory-ladenness.

[7] Duhem's distinction between theory-ladenness in physics and physiology will be revisited.

Thus, left with "theory-ladenness through appeal to causal understanding" (Hanson) and "theory-ladenness through theoretical interpretation" (Duhem and Kuhn), and also having argued that theoretical interpretation goes much further than causal understanding, Heidelberger concludes that it is better to reserve the term ‘theory-ladenness’ to refer only to "genuine theoretical interpretation that transcends causal understanding ..." (ibid., p.146).

Regardless of the extent to which Heidelberger's account can be criticized, productive theory-ladenness, as described in section two above, fits none of the notions of theory-ladenness Heidelberger disambiguates. This is because none of those are about methodological errors or representing phenomena in terms of theories and knowledge that provide the foundations for the instrument used by the researcher. In addition, skeptical arguments of fMRI have focused extensively on the kind of theory-ladenness through theoretical interpretation, which Heidelberger associates with Duhem and Kuhn. Productive theory-ladenness in fMRI is neither about any of the kinds of theory-ladenness Heidelberger defines nor is it involved when skeptics charge that fMRI findings are overly laden with cognitive theories. This is because theories enabling productive theory-ladenness in fMRI are obviously not theories about cognition. One crucial thing to note here is that productive theory-ladenness occurs when fMRI is used to achieve the aims of cognitive neuroscience. That is, it emerges out of the convergence of a technique built on physical and physiological knowledge and a group of researchers’ motivation to investigate the neural bases of cognitive processes. When the skeptics charge fMRI research of pernicious theory-ladenness, they refer to the dangerously circular use of cognitive theories in the interpretation of findings. While, on the other hand, productive theory-ladenness arises out of the theories of magnetic resonance and hemodynamics; it is much less about the cognitive theoretical interpretation of findings and much more so about providing reliable findings in the first place.

The themes of the skeptical accounts of fMRI discussed above remind one of the old and well-known difficulties in psychological science that have been with us since its very beginnings. The point at which they cross paths with cognitive neuroscience is when we ask whether or not neuroimaging data could be used to adjudicate between competing theories of cognition, which reflects a theory-centered approach that has not been very fruitful in expanding the philosophical discussion on neuroimaging. The question about the theoretical import of fMRI findings cannot be addressed without careful scrutiny of the methodological characteristics of this technique. Skepticism about the theoretical significance of neuroimaging is to a great extent supported by arguments claiming that inferences in neuroimaging depend on unreliable procedures. Only an

account that specifically addresses aspects of methodological and inferential procedures in fMRI can provide satisfactory answers to the skeptics. The theory-centered approach has not done this sufficiently to shed any light on what can and what cannot be learned from neuroimaging.

Another reason why the theory-centered approach has not been fruitful is that it overly limits the range of philosophical discussions on neuroimaging. If we approach neuroimaging solely as a potential source of evidence to adjudicate between theories of cognition, we may overlook important questions about the use of instruments and kinds of theory-ladenness in experiment and inference. Instead, when one expands the set of questions about fMRI to include its detailed workings, philosophical discussions on neuroimaging will be rendered richer as well as more easily connected to other philosophical discussions on the use of instruments in other fields of science, which can potentially lead to fruitful results in general philosophy of science. Indeed, Kuhn also has pointed at the significance of the use of instruments in experiments. In *The Structure* he wrote, "At a level lower or more concrete than that of laws and theories, there is, for example, a multitude of commitments to preferred types of instrumentation and to the ways in which accepted instruments may legitimately be employed" (Kuhn, 1962/2012; pp.40-41). In fact, this is related to the distinction between theory-ladenness through theoretical interpretation and theory-ladenness through causal understanding mentioned above. In an experiment, we can talk about theories of the phenomena under study and theories of the instrument used to generate measurements. Furthermore, these theories may be independent of each other. Heidelberger (2003) points to this theme in Kuhn: "... even in Kuhn a sense turns up in which experiment can be independent of the theoretical commitments of a paradigm and dependent only on an entrenched tradition of instrumentation ..." (p.145). In an fMRI experiment, we can talk about theories of cognition and theories of the instrument, i.e. theories of magnetic resonance and hemodynamics, which are independent of theories of cognition. The fMRI scanner is used as a powerful causal agent in generating data, but its workings are not embedded in theories of cognition. As Mayo (1996) and Heidelberger (2003), among others, diagnose, Kuhn's insights into normal science and experiments have not been pursued further by Kuhn or his followers. Nonetheless, the possible occurrences of productive theory-ladenness in other fields of research, especially those that rely on complex measurement and data collection techniques, can be investigated and potentially lead to significant contributions to general philosophy of science. In describing the productive theory-ladenness in fMRI, one of my aims is to contribute to the pursuance of these insights into theory-ladenness and use of instruments in experiment.

4. The Contributions of Productive Theory-Ladenness:

Productive theory-ladenness contributes in two different ways to cognitive neuroscience, namely in the senses (a) and (b) as described in section two above. Here, I illustrate these by giving concrete examples of each.

4.1. Identifying and Controlling for Errors:

In the growing body of philosophical work focusing on methodology in neuroimaging, Roskies (2008, 2010a) is especially helpful in framing productive theory-ladenness. She raises a distinction between the actual versus perceived epistemic status of conclusions and suggests that the perceived epistemic status of neuroimages is higher than their real status. Roskies writes, "determining actual epistemic status will involve a characterization of the inferential steps that mediate between observations and the phenomena they purport to provide information about" (Roskies, 2010a; p.197). She introduces the term *inferential distance* to refer to the totality of these inferential steps; the more the inferential steps the bigger the inferential distance. Per Roskies's diagnosis, the problem is a mismatch between the "actual inferential distance" and the "apparent inferential distance" going from brain activity to the neuroimages presented as findings. Furthermore, she suggests that this inferential distance cannot be univocally characterized.

Roskies is definitely right in saying that neuroimaging results are too often overinterpreted. However, her suggestion that the inferential distance cannot be univocally characterized can be resisted. It is certainly true that there is a great number of technical and inferential procedures in neuroimaging that have to be carried out between initial measurements of brain activation and final brain images. Because of the complexity of these procedures, Roskies suggests that the number and nature of these inferential steps cannot be sufficiently characterized, which lowers the reliability of inferences (Roskies, 2010a). I argue that Roskies's inferential distance problem can be satisfactorily addressed when we break down an experiment into its component parts from design to data collection and analysis and then assess the error probabilities associated with each component. In an fMRI experiment, if these error probabilities are high, then researchers would have to concede that we cannot obtain reliable evidence from the experiment. For example, the scanner used in the experiment may have been oversensitive and detected background noise as task-related activation. If the error probabilities associated with component parts of an experiment are low enough to rule out or minimize errors, then we can safely conclude that researchers can obtain reliable data from the experiment and make warranted inferences. This is how we can go the inferential distance, as it were, and it is made possible thanks to productive theory-ladenness, which provides the theoretical

and technical knowledge of the workings of fMRI, errors that may occur in its use, and ways in which these errors can be formulated and controlled.

I suggest that the dependence of fMRI on these complex procedures is what creates the productive theory-ladenness which enables researchers to represent and study cognitive phenomena in terms of cerebral blood flow. Findings in fMRI experiments are laden with the theories of magnetic resonance and hemodynamics and this reminds one of a distinction Duhem (1906/1991) had raised between physics and physiology. The physiologists make their observations using measurement techniques based on the established theories of physics, whereas physicists have to test their theories based on the theories of physics. That is, physicists do not have the bonus of relying on theories previously established by another field of research. Cognitive neuroscientists using fMRI have to rely on not only physics but also on physiology. At first look, it may seem that cognitive neuroscientists are at a yet higher degree of theory-ladenness than physiologists. However, if we remember that these theoretical bases sit on solid foundations, we can see that this kind of theory-ladenness is productive because it allows for what Hacking (1983) would call representation of and intervention in cognitive phenomena using concepts and methods provided by theories of magnetic resonance and hemodynamics.

The above account can also be related to Hacking's explanation of how the reality of dense bodies in blood is established using different types of experimental techniques, namely, electron microscopes and fluorescent staining, which yield the same result. Specifically speaking, in both types of experiments, small dots in red blood platelets are observed. In order to argue for the reality of these findings, Hacking appeals to an argument from coincidence; it would be a highly improbable coincidence that independent procedures yield the same result unless these small dots are real entities rather than instrumental artifacts. He cites two reasons why the dense bodies are real entities: one reason is the fact that experimental instruments using different physical theories yield the same observations, and the other is that we have a clear understanding of the physical theories that are used to build those instruments. Early in the development of fMRI, experiments were done on cognitive processes of which we have a clear understanding from previous non-imaging research (e.g. perceptual or motor functions) to check for its reliability. These experiments have yielded observations that agreed with the previously established findings. For example, we know from previous biopsychological research that visual perception is related to activation in the brain area known as the occipital cortex. It was shown in experiments, in which subjects performed visual tasks, that the scanner registered high degrees of activation in the same region (e.g. see Kwong et

al., 1992; Ogawa et al., 1992). This is an example of what is stated in proposition (1.2) in section two above. It would be a preposterous coincidence that various types of experiments using different tools and paradigms yield the same kind of artifactual result unless visual functions are indeed related to activation in the occipital cortex.

The stronger reason that fMRI is a reliable tool is that we have a clear understanding of the theories on which it is built. Our scientific understanding of magnetic resonance and hemodynamics gives rise to productive theory-ladenness, which enables researchers to identify and control for the types of errors they may commit using the fMRI scanner, i.e. in the sense (a) of productive theory-ladenness, which also makes possible sense (b) of productive theory-ladenness, both described in section two above. For example, scanners that use high-strength magnetic fields may lead to false positives and this can be scrutinized as an error-characteristic of the scanner. The fMRI scanner works by generating a powerful magnetic field. By applying a strong magnetic field to the chamber inside the scanner, it detects inhomogeneities in the magnetic field as the blood-oxygen-level-dependent (BOLD) response. There are two fundamental problems in this process: the first problem is that the change in the BOLD signal between experimental conditions is very small; what the scanner detects as the difference between conditions is an absolute but very small effect. The ratio of the intensity of the task-related signal and its general variability due to all sources of noise yields a very small value. The second problem has to do with variability in the signal over time, which is influenced by several factors. For example, the temperature of the subject's body, head motion, heart rate, and respiration are all factors that influence the variability of the signal in addition to the cognitive task being performed. Consequently, the task-related change in the BOLD signal is very small when compared to its total variability, so there is a danger of the task-related change in the BOLD signal being masked by other sources of variability. In other words, the signal of interest may easily be lost in the noise. Because of this, some fMRI experiments end up lacking sufficient power to detect task-related signals of interest. Researchers deal with this problem by defining the functional signal-to-noise ratio (SNR); the signal is defined as the difference between two states of brain activity hypothesized to be caused by the cognitive task performed and noise is defined as the overall variability in data over time. The functional SNR is the ratio between these two quantities (for a detailed description, see Huettel et al., 2004, 2008). The higher the functional SNR, the easier it is to detect task-related changes in data. There are several different ways of improving the functional SNR, one of which is to use scanners that generate stronger magnetic fields. The strength of a magnetic field is measured in terms of the *Tesla* (T) unit; as a reference point, the strength of

the earth's magnetic field is 0.00005T. In cognitive neuroscience, scanners with magnetic field strengths from 1.5T to 7.0T are employed (Lazar, 2008; Hashemi et al., 2010). A primary factor determining functional SNR is net magnetization, which is proportional to the strength of the magnetic field, so SNR increases roughly linearly with field strength. Thus, as researchers use scanners of higher magnetic field strength, the functional SNR is improved.

However, one issue with increased magnetic field strength is that at higher field strengths more noise is detected by the scanner in addition to the task-related signal. For example, physiological noise increases quadratically with field strength (Huettel et al., 2004; p.239). Consequently, as field strength is increased, there is a greater danger of the scanner detecting noise as if it is a real effect. When scanners of higher field strengths are used, task-related changes in the BOLD signal may be lost in increased physiological noise. Thus, at field strengths higher than 4.0T we may get too many cases where noise is detected as a real effect. This has been noted by Savoy (2001) as a concern about fMRI in general; he also argues that this is similar to the problem raised by Meehl (1967) who claimed that if we sufficiently increase the power of significance tests, we can reject any null hypothesis even if it is exactly true. Likewise, it appears that if we use a high-strength scanner, we may, with greater probability, find supporting evidence for hypotheses about activation-task performance mappings even if they are false. The observed activation may simply be noise and have nothing to do with the cognitive task. This is an error characteristic of a certain component of the experiment, which must be scrutinized. In general, one could argue that increasing field strength increases the probability of errors and researchers can use this knowledge as a guide in evaluating the possibility that they have a real effect as opposed to an artifact of the scanner.

Naturally, there are other factors that may also lead to the detection of noise as a real effect or other types of errors, e.g. preprocessing protocols, multiple testing, thresholding, or spatial or temporal filtering problems. Potential errors stemming from these factors can be similarly identified, formulated, and minimized thanks to productive theory-ladenness.[8] For example, several groups of researchers have conducted a series of studies assessing different preprocessing protocols, which gave rise to the "nonparametric prediction, activation, influence, and reproducibility

[8] Having proper error probabilities between 0 and 1, as formulated in Mayo's error-statistical account would be a more effective way of controlling for potential errors, for details see Mayo (1996) and Mayo & Spanos (2011). These error probabilities can be calculated in a series of studies similar to those in which false positive and false negative rates of diagnostic tests are calculated.

resampling" or NPAIRS framework (Strother et al., 2002; LaConte et al., 2003). This framework applies cross-validation: an fMRI data set is split into two halves, one half is designated as the "training" data and used to estimate the parameters for a predetermined model. The estimated parameters and the model are used to make predictions to be tested on the other half of the data designated as "test" data. This process is repeated but with the training and test data switched, i.e. in the second application of the process test data are used for training and vice versa. In several runs, reproducibility of the experimental findings is assessed by comparing the results of statistical analyses on both halves of the data. The flexible nature of this framework allows researchers to assess the effects of different types of preprocessing protocols.

In one study, LaConte et al. (2003) applied different protocols, called analysis chains, which included different levels of preprocessing of raw fMRI data. One chain included no preprocessing procedures, whereas others included normalization and different degrees of spatial filtering, i.e. one chain applied a narrow filter and another chain applied a wide filter. Then, they conducted statistical analyses on data sets that came from these different analysis chains in order to assess the effects of different preprocessing protocols on prediction accuracy and reproducibility. The results showed that spatial filtering (smoothing) was the most effective procedure for improving prediction accuracy and reproducibility. However, as LaConte and his colleagues note, there are no general pre-data guidelines for the optimal preprocessing protocol for all experiments (ibid.). The optimality of a preprocessing protocol is dependent not only on the elements of the protocol, as in how much smoothing or normalization was applied, but also on other experimental parameters such as the type of scanner used, design of the experiment, etc. As Lazar (2008) and LaConte et al. (2003) suggest, researchers can apply different preprocessing protocols to the same set of raw data, then do statistical analyses on the data sets that the different protocols yield. In this way, they can assess the effects of these protocols on the same data set. As researchers become more aware of the errors that preprocessing procedures may introduce, they can start devising methods of identifying and controlling for these errors. The NPAIRS framework is one example, among many, illustrating how productive theory-ladenness enables researchers to improve the reliability of methodologies and inferences (in sense (a) as described above).

4.2. Representing and Investigating Cognitive Phenomena:

In the previous section, we have seen how productive theory-ladenness enables fMRI researchers to identify, and control for, methodological and inferential errors. I have also proposed that productive theory-ladenness in sense (a) also makes possible sense (b) of productive theory-ladenness as

described in section two. That is, it enables researchers to represent and investigate cognitive phenomena in terms of hemodynamic data as subjects perform well-defined cognitive tasks. This can be illustrated by looking at the development of the HERA model of memory on the basis of neuroimaging findings.

The hemispheric encoding/retrieval asymmetry (HERA) model was proposed by Tulving and his colleagues in 1994 (Tulving et al., 1994). When it was first introduced, the HERA model was a straightforward description of empirical regularities obtained in PET studies of memory—although it works differently from fMRI, PET also provides measurements of cerebral blood flow.[9] Researchers had obtained differential activation patterns in left and right prefrontal cortical regions when subjects engaged in encoding and retrieval tasks of episodic memories. In the context of HERA, the terms 'semantic memory' and 'episodic memory' can be used simply to refer to the different kinds of experimental tasks without having to worry too much about the theoretical baggage they may carry. These terms denote common phenomena of memory in everyone's experience, such as remembering something or learning something new, and they need not mean anything more theoretically complex than that in the context of this discussion.[10]

In 1994, when the HERA model was proposed, cognitive neuroscience was just becoming a discipline of its own and researchers were beginning to obtain new empirical regularities. In this environment, HERA was proposed in a data-driven manner on the basis of PET findings. Tulving and colleagues (1994) did a series of PET studies and also reviewed studies from different laboratories in which subjects performed three types of tasks; semantic memory retrieval, episodic memory encoding, and episodic memory retrieval. They saw some regularities in the observed patterns of brain activation and summarized these regularities in a set of hemodynamic statements: 1) Left prefrontal cortical regions are activated in semantic memory retrieval to a greater extent than right prefrontal cortical regions; 2) Left prefrontal cortical regions are activated in encoding novel features of retrieved information into episodic memory to a greater extent than the right prefrontal

[9] Though on the basis of different physical knowledge, productive theory-ladenness occurs in PET, too. Indeed, it should be clear by now that productive theory-ladenness occurs in any successful use of a complex instrument of measurement or observation.

[10] Nonetheless, it is still the case that experimental knowledge coming from behavioral experiments in the cognitive psychology of human memory played a major role in designing the tasks used in neuroimaging experiments. Thus, the question arises; 'to what extent does experimental knowledge from cognitive psychology provide background knowledge for neuroimaging experiments?' This question is related to issues of cognitive ontology, which I plan to address elsewhere, but it is not directly relevant to my purposes here.

cortical regions; and, 3) Right prefrontal cortical regions are involved in episodic memory retrieval to a greater extent than left prefrontal cortical regions. Essentially, these three hemodynamic statements constituted the initial version of the model.

In 1996, Nyberg, Cabeza, and Tulving published a review article reporting results from both PET and fMRI experiments. Again, the great majority of these experiments exhibited the same findings predicted by HERA with only a few exceptions (Nyberg, Cabeza, & Tulving, 1996). Gabrieli et al. (1998) used similar encoding tasks in which subjects were shown pictures and found a significantly high degree of activation in the right inferior frontal cortical regions as subjects performed encoding tasks. In another study, Kelley et al. (1998) found a significantly high degree of activation in the right prefrontal cortical regions in a similar encoding task. Both sets of findings contradict the model, which predicts a higher degree of activation in the left prefrontal cortex in encoding tasks. Some researchers concluded on the basis of these findings that the asymmetry of prefrontal cortex activations is due to the type of stimuli used, i.e. verbal versus non-verbal, rather than being caused by encoding versus retrieval.

Unsatisfied with their explanations of the findings contradicting HERA, Habib, Nyberg, and Tulving (2003) reformulated the model to be stricter and more precise in its assertions. But also, they insisted that HERA was a set of statistical hypotheses that compared degrees of activation in the prefrontal cortical regions between encoding and retrieval tasks rather than a set of absolute hypotheses. In this reformulation, there were two specific hemodynamic hypotheses which were expressed using abbreviations: ‘Enc’=encoding, ‘Ret’=retrieval, ‘L’=a given left prefrontal cortical region and ‘R’=corresponding region in the right prefrontal cortex. Combinations of task (Enc or Ret) and regions (L or R) stood for the observed activation during a given task in a given region, e.g., ‘Enc L’ stood for the activation observed in a specific region in the left prefrontal cortex during an encoding task. Thus, the two hemodynamic hypotheses that constitute the model were stated: 1. (Enc L–Ret L) > (Enc R–Ret R); and 2. (Ret R–Enc R) > (Ret L–Enc L) (ibid., p. 241). Several fMRI experiments, as well as experiments that used newer neuroimaging techniques, yielded results that supported the model since its reformulation (see Babiloni et al., 2006, Cole 2006, Thimm et al., 2010, and Okamoto et al., 2011).

HERA was proposed as a description of a set of hemodynamic findings showing asymmetry in patterns of brain activation between encoding and retrieval tasks in episodic memory. It is to be noted that strict interpretations of modularist theories imply that there cannot be a single model like HERA, because, for some of these theories, neuroimaging results make sense only if they talk about

cognitive modules executing specific computations defined by the theory and located in a neuroanatomically distinct part of the brain. This is why some suggest that before we can use neuroimaging techniques to study cognition we first have to have an accurate and complete psychological theory of cognitive systems down to the most specific, separate function. Yet, all the experiments done or cited by Tulving and his colleagues in support of HERA reported activations from several cortical regions. These regions were all in the prefrontal cortex but were either adjacent to each other or a few centimeters apart. For some modularist theories, it is not clear whether or not these different regions of the prefrontal cortex would be taken to be neuroanatomically distinct to a sufficient degree. Thus, HERA will not make much sense for anyone who adheres to a strictly modularist theory.

For some others, these findings may constitute a refutation of strictly modularist theories of cognition. A defense of neuroimaging against putting rigid theoretical requirements on what can be accepted as a genuine finding has been offered by Roskies (2010b). Roskies has also rightly argued that neuroimaging never truly claimed to support such strong theoretical conclusions as strict modularity or the like. Instead, it has the more pragmatic goal of revealing what it can about regularities between neural phenomena and cognitive function (ibid.). What I wish to argue is that independently of what large-scale cognitive theory one adopts, the HERA model stands as a set of robust hemodynamic findings and it is fruitful for cognitive neuroscience to talk about a well-established empirical regularity in observed brain activation patterns in the prefrontal cortex during encoding and retrieval tasks. In addition, HERA came out of a data-driven approach and was described mostly in terms of hemodynamic hypotheses, which did not assume too much about cognitive theories of human memory systems. For example, even if we stop using cognitive theoretic terms such as semantic or episodic memory, we can still talk about the model in terms of specific, well-defined memory tasks. The hemodynamic findings, on the basis of which HERA was developed, would still stand when theories of cognitive science change and new cognitive ontologies divide and categorize human memory differently, or exclude altogether strict categorizations, or redefine cognition to include processes and entities outside the brain. Come what may theoretically, cognitive scientists will still have to accommodate the HERA findings in terms of whatever cognitive constructs they adopt.

After it was initially proposed, several groups of researchers have done various neuroimaging experiments to test the HERA model, a great majority of which yielded supporting results. Also, however, it was reformulated to be more precise in its assertions in response to some

seemingly contradictory results. Now, HERA is a well-established model, built on hemodynamic findings, that describes a certain pattern in which the human brain works when individuals engage in memory tasks of encoding and retrieval. This model is a contribution of cognitive neuroscience to the general understanding of relationships between the human brain and cognition. Researchers were able to formulate, test, and refine HERA using neuroimaging relatively independently of large-scale theories of cognition and this was made possible, at least partially, thanks to productive theory-ladenness.

5. Conclusion:

I have proposed that the scientific foundations of fMRI give rise to productive theory-ladenness in cognitive neuroscience. I have also proposed that there are two ways in which this productive theory-ladenness occurs, namely, it enables researchers to (a) identify and control for the types of methodological and inferential errors that may arise in fMRI studies, and, (b) represent and investigate neural and cognitive phenomena in terms of hemodynamic data as subjects perform well-defined tasks. We have seen how these two aspects of productive theory-ladenness can be illustrated with concrete examples; such as correcting for physiological noise and identifying the most reliable preprocessing protocol using the NPAIRS framework as examples of (a) and the discovery and development of the HERA model as an example of (b). These two aspects also reflect the two different paths of growth of experimental knowledge in cognitive neuroscience, namely, 1) the growth of knowledge improving the reliability of measurement and data analyses, and 2) the growth of knowledge of connections between patterns of neural activation and performance of cognitive tasks. The development and use of NPAIRS illustrate (1) and the evolution of HERA, as new neuroimaging data became available, illustrates (2).

The arguments and examples above motivate a more pragmatic and fruitful approach in the philosophy of neuroimaging. When we look at fMRI with an eye toward appreciating the kinds of knowledge it can reliably provide, we can see it as a tool for expanding our knowledge on relationships between hemodynamic processes and cognition instead of limiting it as only a novel tool for adjudicating between existing theories of cognition. When fMRI, and other similar instruments, are used in methodologically robust ways, the results can be used not only to adjudicate between existing theories but also can yield novel findings and insights into relationships between neural and cognitive phenomena independently of these theories. To use Hacking's famous phrase, experiment in cognitive neuroscience has “a life of its own,” which has two related paths of growth made possible thanks to productive theory-ladenness.

References:

Aktunc, M. Emrah. 2011. "*Experimental Knowledge in Cognitive Neuroscience: Evidence, Errors, and Inference*." PhD diss., Virginia Polytechnic Institute and State University.

Babiloni, C., Vecchio, F., Cappa, S., Pasqualetti, P., Rossi, S., Miniussi, C., Rossini, P.M. (2006). "Functional Frontoparietal Connectivity During Encoding and Retrieval Processes Follows HERA Model: A High-Resolution Study." Brain Research Bulletin, 68, 203 – 212.

Baddeley, A.D. & Hitch, G. (1974). "Working Memory." In G. H. Bower (ed.), *The Psychology of Learning and Motivation*, Vol. 8. London, UK: Academic Press.

Baddeley, A.D. (2003). " Working Memory and Language: An Overview" *Journal of Communication Disorders, 36* (3), 189-208.

Bechtel, W. (2002a). "Decomposing the Mind-Brain: A Long-Term Pursuit." *Brain and Mind, 3*, 229-242.

Bechtel, W. (2002b). "Aligning multiple research techniques in cognitive neuroscience: Why is it important?" *Philosophy of Science, 69*, S48–S58.

Chang, H. (2012). *Is Water H_2O?: Evidence, Realism and Pluralism.* London, UK: Springer.

Cole, M.A. (2006). "Effects of Goal-Setting on Memory Performance in Young and Older Adults: A Functional Magnetic Resonance Imaging (fMRI) Study." *Dissertation Abstracts International: Section B: The Sciences and Engineering*, 5134.

Coltheart, M. (2006). "What Has Functional Neuroimaging Told Us About The Mind (so Far)?" *Cortex, 42*, 323-331.

Coltheart, M. (2010). "What Is Functional Neuroimaging For?" In Hanson, S. J. & M. Bunzl (eds.) *Foundational Issues in Human Brain Mapping*. Cambridge, MA: The MIT Press.

Duhem, P. (1906/1991). *The Aim And Structure of Physical Theory.* [Translated from the French by Philip P. Wiener] Princeton, NJ: Princeton University Press.

Gabrieli, J.D.E., Poldrack, R.A., & Desmond, J.E. (1998). "The Role of Left Prefrontal Cortex in Language and Memory." *Proceedings of the National Academy of Sciences, 95*, 906 – 913.

Habib, R., Nyberg, L., & Tulving, E. (2003). "Hemispheric Asymmetries of Memory: The HERA Model Revisited." *Trends in Cognitive Sciences, 7*, 241 – 245.

Hacking, I. (1983). *Representing and Intervening: Introductory Topics in the Philosophy Of Natural Science*. Cambridge, UK: Cambridge University Press.

Hanson, N.R. (1958). *Patterns of Discovery*. Cambridge, UK: Cambridge University Press.

Hardcastle, V.G. & Stewart, C.M. (2002). "What Do Brain Data Really Show?" *Philosophy of*

Science, 69, S72-S82.

Hashemi, R.H., Bradley, Jr., W.G., & Lisanti, C.J. (2010). *MRI: The Basics*, Third Edition. Philadelphia, PA: Lippincott Williams & Wilkins.

Heidelberger, M. (2003). "Theory-Ladenness and Scientific Instruments in Experimentation" In H. Radder, Ed. *The Philosophy of Scientific Experimentation,* pp.138-151. Pittsburgh, PA: University of Pittsburgh Press.

Henson, R. (2005). "What Can Functional Neuroimaging Tell The Experimental Psychologist." *The Quarterly Journal of Experimental Psychology*, *58A*, 193-233.

Huettel, S.A., Song, A.W., & McCarthy, G. (2004). *Functional Magnetic Resonance Imaging*. Sunderland, MA: Sinauer Associates, Inc. Publishers.

Huettel, S.A., Song, A.W., & McCarthy, G. (2008). *Functional Magnetic Resonance Imaging, 2nd Edition*. Sunderland, MA: Sinauer Associates, Inc. Publishers.

James, W. (1890). *The Principles of Psychology*. New York, NY: Henry Holt and Co.

Kelley, W.L., Miezin, F.M., McDermott, K., Buckner, R.L., Raichle, M.E., Cohen, N.J., & Petersen, S.E. (1998). "Hemispheric Specialization in Human Dorsal Frontal Cortex and Medial Temporal Lobes for Verbal and Nonverbal Memory Encoding." *Neuron, 20*, 927 – 936.

Klein, C. (2010). "Philosophical Issues in Neuroimaging." *Philosophy Compass, 5*, 186-198.

Kuhn, T. (1962/2012). *The Structure of Scientific Revolutions* (Fourth Ed.). Chicago, IL: The University of Chicago Press.

Kwong, K.K., Belliveau, J.W., Chesler, D.A., Goldberg, I.E., Weisskoff, R.M., Poncelet, P.B., Kennedy, D.N., Hoppel, B.E., Cohen, M.S., Turner, R. (1992). "Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation." *Proc. Natl. Acad. Sci.*, *89* (12), 5675-5679.

LaConte, S., Anderson, J., Muley, S., Ashe, J., Frutiger, S., Rehm, K., Hansen, L.K., Yacoub, E., Hu, X., Rottenberg, D., & Strother, S. (2003). "The Evaluation of Preprocessing Choices in Single-Subject BOLD fMRI Using NPAIRS Performance Metrics." *NeuroImage, 18*, 10–27.

Lazar, N. A. (2008). The Statistical Analysis of Functional MRI Data. New York, NY: Springer.

Le Bihan, D. (2014). *Looking Inside The Brain: The Power of Neuroimaging*. Princeton, NJ: Princeton University Press.

MacCorquodale, K. & Meehl, P.E. (1948). "On A Distinction Between Hypothetical Constructs and Intervening Variables." *Psychological Review*, *55*, 95-107.

Mayo, D. (1996). *Error and the Growth of Experimental Knowledge.* Chicago, IL: The University

of Chicago Press.

Mayo, D. & Spanos, A. (2011). “Error Statistics.” In Prasanta S. Bandyopadhyay and Malcolm Forster (eds.) *The Handbook of Philosophy of Science, Volume7: Philosophy of Statistics*. Amsterdam, The Netherlands: Elsevier Publishers.

Meehl, P.E. (1967). “Theory-testing in Psychology and Physics: A Methodological Paradox.” *Philosophy of Science, 34,* 103-115.

Nathan, M.J. & Del Pinal, G. (2017). “The Future of Cognitive Neuroscience? Reverse Inference in Focus.” *Philosophy Compass, 12*: e12427.

Nyberg, L., Cabeza, R., & Tulving, E. (1996). “PET Studies of Encoding and Retrieval: The HERA Model.” *Psychonomic Bulletin and Review, 3,* 135 – 148.

Ogawa, S., Tank, D.W., Menon, R., Ellermann, J.M., Kim, S.G., Merkle, H., Ugurbil, K. (1992). "Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging." *Proc. Natl. Acad. Sci., 89* (13), 5951-5955.

Okamoto, M., Wada, Y., Yamaguchi, Y., Kyutoku, Y., Clowney, L., Singh, A.K., Dan, I. (2011). “Process-specific Prefrontal Contributions to Episodic Encoding and Retrieval of Tastes: A Functional NIRS Study.” *NeuroImage, 54,* 1578 – 1588.

Roskies, A. (2008). “Neuroimaging and Inferential Distance.” *Neuroethics, 1*, 19-30.

Roskies, A. (2009). “Brain-Mind and Structure-Function Relationships: A Methodological Response to Coltheart.” Philosophy of Science, 76, 927-939.

Roskies, A. (2010a). “Neuroimaging and Inferential Distance.” In Hanson, S. J. & M. Bunzl (eds.) *Foundational Issues in Human Brain Mapping*. Cambridge, MA: The MIT Press.

Roskies, A. (2010b). “Saving Subtraction: A Reply to Van Orden and Paap.” *British Journal for the Philosophy of Science, 61,* 635-665.

Savoy, R.L. (2001). “History and Future Directions of Human Brain Mapping and Functional Neuroimaging.” Acta Psychologica, 107, 9-42.

Strother, S.C., Anderson, J., Hansen, L.K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., & Rottenberg, D. (2002). “The Quantitative Evaluation of Functional Neuroimaging Experiments: The NPAIRS Data Analysis Framework.” *NeuroImage, 15,* 747–771.

Thimm, M., Krug, A., Markov, V., Krach, S., Jansen, A., Zerres, K., Eggermann, T., Stocker, T., Shah, N.J., Nothen, M.M., Rietschel, M., & Kircher, T. (2010). “The Impact of Dystrobrevin-Binding Protein I (DTNBPI) on Neural Correlates of Episodic Memory Encoding and

Retrieval." *Human Brain Mapping, 31*, 203 – 209.

Tulving, E., Kapur, S., Craik, F.I.M., Moscovitch, M., Houle, S. (1994). "Hemispheric Encoding/Retrieval Asymmetry in Episodic Memory: Positron Emission Tomography Findings." *Proceedings of the National Academy of Sciences, 91*, 2016 – 2020.

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

Uttal, W. (2002a). "Functional Brain Mapping – What Is It Good For? Plenty, but not Everything! (A Reply to Malcolm J. Avison)." *Brain and Mind*, 3, 375-379.

Uttal, W. (2002b). "Précis of The New Phrenology: The Limits of Localizing Cognitive Processes in the Brain." *Brain and Mind*, 3, 221-228.