Tackling Duhemian Problems in Neuroimaging: An Alternative to Skeptical Approaches in Philosophy of Cognitive Science

M. Emrah Aktunc

Ozyegin University

Abstract: Duhem's problem arises in different fields of science, especially in contexts where the tools and procedures of measurement and analysis are numerous and highly complex. Several philosophers of cognitive science, as well as cognitive scientists, have pointed to its manifestations in fMRI as grounds for skepticism regarding the epistemic value of neuroimaging results. I offer an alternative approach to neuroimaging, based on Deborah Mayo's error-statistical account, to address Duhemian arguments for skepticism of neuroimaging in philosophy of cognitive science. Duhem's problem in fMRI is more fruitfully approached in terms of error probabilities as formulated by Mayo. This is illustrated in examples such as the use of probabilistic brain atlases, comparison of different preprocessing protocols with respect to their error characteristics, and statistical modeling of fMRI data. These examples demonstrate the ways in which we can better understand and formulate the general methodological problem and direct the way toward more balanced approaches to neuroimaging in philosophy of cognitive science that will more accurately identify what to be skeptical about and what epistemic contribution neuroimaging can reliably provide.

Duhem's problem arises when a scientist does an experiment, or a series of experiments, to test some hypothesis H and gets a result that does not agree with the hypothesis. One construal of the problem is to think of it in terms of a *modus tollens* of the type Popper discussed: If hypothesis H, then data *e*. Not-*e*. Therefore, not-H. Of course, in actual scientific practice, things do not work this way. It is rather like this: If H_1, H_2, H_3,…, H_n and A_1, A_2, A_3,…,A_n, then *e*. Not-*e*. Therefore, not-H_1 or not-H_2 …, or not-A_1 or not-A_2… where H_1 through H_n and A_1 through A_n are auxiliary hypotheses and assumptions involved in the experiment that yielded not-*e*. The latter inference is the only one that deductively follows. Thus, we do not know if it is the hypothesis that we should blame for

not-*e* and falsify H, or we should hang on to H as it may be any of the auxiliary hypotheses or assumptions that are responsible for obtaining not-*e*. Several solutions to Duhem's problem have been proposed by philosophers of science. In this paper, I will first discuss how Duhem's problem manifests itself in functional neuroimaging, looking at fMRI as a representative neuroimaging medium, and then propose how it may be addressed employing the error-statistical notions of severe tests and error probabilities.

The goal in an fMRI experiment is to relate changes in brain physiology over time to an experimental manipulation (Huettel et al., 2008). One essential type of inference is about where in the brain, if anywhere, there is significant activity measured by the fMRI scanner as participants perform a cognitive task of interest compared to a control condition in which they do nothing or do a simple task. This kind of inference is mostly drawn across participants; it can take the form "participants had significant activity in brain region X as they performed cognitive task C." In fMRI experiments, this kind of inference is usually embedded into a statistical model and linked to the alternative hypothesis in a Neyman-Pearsonian significance test. This alternative hypothesis can be framed in terms of parameters μ_0 and μ_1 with μ_0 designating mean activation in a certain brain region X in the control condition and μ_1 designating mean activation in the same region X in the experimental condition in which participants perform the cognitive task of interest. The alternative hypothesis takes the form H_a: $\mu_1 - \mu_0 > 0$ to be tested against the null hypothesis H_0: $\mu_1 - \mu_0 = 0$ in a significance test formulated in the context of a statistical model of the fMRI data. At first glance, this seems to be unproblematic assuming that the many aspects of the experiment, such as statistical modeling and analyses, were free of biases and/or misspecification. Later I will get back to issues of statistical models of fMRI data, but there are other difficulties that arise in fMRI before the stage of statistical modeling and significance tests. Neuroimaging researchers, e.g. Huettel et al. (2004; 2008), point to certain issues about inferences to mappings between patterns of neural activity and specific brain regions. These issues stem from the difficulty of satisfactorily addressing questions such as 'how do neural activity map onto neuroanatomy?', 'how consistent is that mapping across participants?', or 'how do functional data "correspond" to underlying neuroanatomy?'

To address these questions, fMRI data have to be mapped onto high resolution structural images. However, we have to remember the fact that people's brains vary with respect to size, shape, orientation, and gyral anatomy. Brain sizes of two participants in a given fMRI experiment may differ by 30 per cent. A hidden assumption in data analyses is that in each voxel (volumetric pixel) the fMRI scanner represents a unique and unchanging location in the brain. Given neuroanatomical variability, this assumption is always wrong. For example, voxel M may correspond to region X in one participant while the same voxel may correspond to region Y in another participant. Brain shapes of individuals differ a great deal, as in long and thin versus short and fat brains. The organization of sulci and gyri is also variable across individuals in ways that major landmarks in the brain may be at different positions and differently oriented across individuals. Because of neuroanatomical variability, when we draw an inference of the form "participants had significant activity in brain region X as they performed cognitive task C" we do not know whether or not in each participant the activity was really in region X. For a given participant B, it may be true that there is significant activity in his/her brain, but it may be in region Y adjacent to region X. That is, what corresponds to region X according to the mapping used by the fMRI scanner may in fact be region Y in participant B's brain. This is problematic for any generalizations that associate a brain region with a given cognitive process. For example, let us say that in an experiment participants were asked to perform cognitive task C and the results show that they had significant activation in brain region X. We conclude that region X is involved in the performance of cognitive task C. However, participant B, whose brain anatomy differs from other participants, performed the same cognitive task but the activation may have been in region Y of her brain. While we may assume that she had significant activation in region X, we do not know if it really was region X or region Y that was activated, because we did not take into account neuroanatomical variability. Of course, neuroanatomical variability in other participants' brains may likewise complicate our inferences. Therefore, our generalization may have been in error and we would not know if we committed this error or how probable it was that we committed this error in this experiment.

To address this problem, researchers apply a procedure called normalization in

which shape differences across brains are compensated for by mathematically stretching, squeezing, and warping each brain so that it is the same as other brains. In most normalization procedures the Talairach stereotaxic space is used, which is a coordinate system of the brain that defines locations of brain structures in terms of their coordinates (Talairach & Tournoux, 1988). The actual brain that was used by Talairach and Tournoux to develop this system was that of an elderly lady. This creates problems of representativeness, because participants in fMRI experiments would probably have brains that are different from the brain that is taken as a model by the Talairach space. Nonetheless, the probability of drawing false inferences may be reduced to some extent by normalization. However, since we do not have empirical measures of the variability across brains and the representativeness of the brain used in the Talairach space is questionable, we cannot safely assume that errors due to neuroanatomical variability are sufficiently reduced. In other words, normalization based on Talairach space does not give us the accurate, or even approximate, error probabilities associated with inferences of the form "participants had significant activity in brain region X as they performed cognitive task C."

Let us state the problem in Duhemian terms. In this context, our experimental hypothesis may be stated thus: H: Brain region X is involved in cognitive task C. Then, the *modus tollens* will be the following: If H: brain region X is involved in cognitive task C, then *e*: as participants perform C, there will be a significantly higher level of activity in X compared to the control condition where participants do nothing or perform a simpler or different task. As described above, this experimental hypothesis is linked to the statistical alternative hypothesis H_a: $\mu_1 - \mu_0 > 0$ to be tested against the null hypothesis H_0: $\mu_1 - \mu_0 = 0$ in a significance test formulated in the context of a statistical model of the fMRI data. Let us assume we carry out the experiment and we do not observe significantly higher activity in X, so we get not-*e*. With this result, we do not reject the null hypothesis. Therefore, we conclude not-H; brain region X is not involved in the performance of cognitive task C. To this conclusion we can object with the Duhemian argument that in actual scientific practice, especially fMRI, experiments are highly complex in their several components and the inferential procedure is rather like this: If H_1, H_2, H_3,…, H_n and A_1, A_2, A_3,…,A_n, then *e*. Not-*e*. Therefore, not-H_1 or not-H_2 … or not- H_n, or not-A_1 or not-A_2… or not- A_n where

H_1 through H_n and A_1 through A_n are auxiliary hypotheses and assumptions of the experiment that yielded not-*e*. H_1 would be the hypothesis above; namely "brain region X is involved in cognitive task C" and *e* would be the prediction "as participants perform C, there will be significant activity in X." Put simply, the normalization procedure is going to be one of the auxiliary assumptions, say A_1, and according to this assumption, normalization takes care of any neuroanatomical variability across participants endangering the reliability of inferences. If not-*e* is obtained, then one could blame A_1 for obtaining not-*e*; that is, one could say that the normalization procedure was not sufficiently effective. Perhaps there was a significant anatomical mismatch between the mapping used by the fMRI scanner and participants' brains. If we remember the shortcomings of the normalization procedure based on the Talairach space, one could easily, and rightly, raise this objection and suggest that the blame for obtaining not-*e* should be put on the ineffectiveness of the normalization procedure and not on the falsity of our experimental hypothesis.

Similar objections can be raised about other aspects of an fMRI experiment, for example we could say that the fMRI scanner was not sensitive enough to detect activity or that the procedure for increasing the signal to noise ratio was not effective. In fMRI, data are collected as a time series, a large amount of data on the hemodynamic processes in the participant's brain are acquired in temporal order at a specified rate as the subject performs a cognitive task. Each session consists of multiple runs of presentation of the cognitive task and each run includes single images of the brain called volumes. Volumes consist of images of slices of the brain and slices consist of three-dimensional voxels. A matrix of voxels makes up the slice where the matrix may be of size 64x64 or 128x128. In an experiment that studies the entire brain there may be as many as 25 slices. For example, in an experiment where the size of the voxel matrix is 64x64 and there are 25 slices in the volume, there would be time series data from a total of 102,400 voxels to be processed and analyzed. The fMRI data set can be thought of as a four-dimensional matrix; voxels by voxels by slice by time. In a simple 6-minute run of an experiment that covers the entire brain and where the fMRI scanner delivers an excitation pulse every second, the four-dimensional matrix of data be 64x64x25x360, where 64x64 is the size of the voxel matrix,

25 is the number of slices, and 360 is the number of volumes since data from the entire brain are recorded every second (Huettel et al., 2004; pp.186-188). Because of the complexity of the fMRI experiment as a whole and the massive size of raw data sets, several computational procedures, collectively called "preprocessing," are needed to obtain data sets in canonical form so that statistical tests can be carried out on the data.

Due to the multiplicity of experimental procedures and inferential steps in fMRI studies, it is extremely easy to raise Duhemian objections when we obtain results that disagree with our experimental hypothesis. Of course, such objections can be raised, too, for experiments the results of which agree with the hypotheses tested. Hence, several philosophers of science have put forth skeptical arguments on the basis of this general Duhemian problem in neuroimaging. One common theme in these arguments is that this problem lowers the reliability of inferences and renders ambiguous the findings in neuroimaging experiments. Bogen (2010) has argued that the dependence of fMRI on complex inferential procedures calls into question the reliability of fMRI as an observational tool, because it is difficult in this kind of experiment to pinpoint what exactly is observed. He writes that fMRI is a type of science where "evidence is produced by processes so convoluted that it's hard to decide what, if anything has been observed" (ibid., p. 11). In such a construal of fMRI methodology, the Duhemian difficulty seems to be taken to an extreme. According to Bogen, the problem seems to be even more difficult than pinpointing blame for a negative result, because he suggests that, in an fMRI experiment, we cannot even be sure what the result is, if there is any.

Adina Roskies is another philosopher of science who has emphasized the methodological complexity of neuroimaging and the problems it creates. She employs a distinction between the actual versus perceived epistemic status of conclusions and suggests that the perceived epistemic status of neuroimages, i.e. the form in which neuroimaging findings are presented, is higher than their real status (2008; 2010). Of course, in order to interpret results correctly, we need to determine the actual epistemic status. Roskies claims that "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. This characterization will include

both the nature of the steps, and their relative certainty…" (Roskies, 2010; p.197). Roskies introduces the term *inferential distance* to refer to the totality of these inferential steps; the more the inferential steps the bigger the inferential distance. In the fMRI literature, some of these steps are referred to as preprocessing of data, but statistical modeling and analysis of data would also be included in what she calls inferential steps.

As Roskies's diagnosis goes, the problem in neuroimaging is the mismatch between the "actual inferential distance" and the "apparent inferential distance" between actual brain activity and the neuroimages that are presented as findings. She writes; "I use 'actual inferential distance' to refer to the inferences explicitly employed in a scientific practice, while 'apparent inferential distance' indicates a more subjective measure characterizing the confidence people place in a conclusion on the basis of evidence" (ibid.). Roskies is definitely right in suggesting that the tendency to overinterpret fMRI results may lead to erroneous conclusions. However, her further assumption that inferential distance in fMRI cannot be univocally characterized can be questioned. It is definitely a fact that there are a great number of technical and inferential procedures in fMRI experiments that have to be carried out between initial measurements of brain activation and neuroimages. These inferential steps require complicated computational procedures on immensely large data sets and, because of the complexity of these procedures, Roskies says that the number and nature of these inferential steps cannot be sufficiently characterized. This, she suggests, lowers the reliability of inferences drawn from data, which leads her to a pessimistic view about the epistemic value of fMRI findings.

How can we address these Duhemian problems in neuroimaging? I wish to approach this question from the perspective of Mayo's error-statistical account (Mayo, 1996; 2005), which offers the kind of conceptual machinery for dealing with the complex nature of functional neuroimaging. The crucial point is that Roskies's inferential distance problem can be satisfactorily addressed when we break down an fMRI inquiry into its component parts. Essentially, these components include design of experiments, collection and preprocessing of behavioral and neuroimaging data, and statistical modeling and inferential procedures such as correlation analyses and hypothesis tests. As required by the error-statistical account, we can assess in a piece-meal fashion the error characteristics

associated with each component. A general problem of inference arising in functional neuroimaging is the difficulty and/or lack of assessments of error probabilities, or error characteristics, of the component procedures employed in experiments. The Duhemian problem can manifest itself in any of these procedures and it has to be addressed at each stage it arises.

The specific issue of inferences of the form "participants had significant activity in brain region X as they performed cognitive task C" is just one of these stages where the problem of assessing error probabilities arises. This problem arises, because, in many experiments, we do not have empirical measures of the variability across participants' brains. We do not know how probable it is to misidentify brain regions to be paired with observed brain activity. This problem can be formulated in terms of Mayo's error-statistical account, according to which the error probabilities associated with the experimental procedure, instrument, or test, are needed in order to assess whether or not the experimental result constitutes good evidence for the hypothesis tested in that experiment. If these error probabilities are difficult to assess and/or not assessed at all, then we have a problem regarding the inference we may wish to draw on the basis of the experimental result. This is precisely the issue in the specific kind of error stemming from neuroanatomical variability.

The problem can also be described in terms of the error-statistical notion of severe tests. Mayo's severity principle states: "Data **x** (produced by process *G*) provide a good indication or evidence for hypothesis *H* (just) to the extent that test *T* severely passes *H* with **x**." (Mayo, 2005; p. 100). When does a hypothesis *H* pass a test *T* severely with data **x**? For this, *two things* must obtain; *first*, data **x** fits or agrees with *H*, and *second*, test *T* would have produced, with high probability, data that fit less well with *H* than **x** does, were *H* false (Mayo, 1996; 2005). The idea here is that data **x** is evidence for hypothesis *H* just to the extent that the accordance between **x** and *H* would be difficult to achieve were *H* false. In other words, one must have done a good job at probing the ways in which one may be wrong in inferring from an accordance between data **x** and hypothesis *H* that *H* is true (or is well tested or corroborated). Here, a very important point to note is that the severity of a test is <u>not</u> a feature of only the test itself. Rather, it is a function of a group of

things; namely, the *test*, (or the experiment, broadly defined as the procedures that generated the data), the *data*, and the specific *hypothesis* about which an inference is drawn (Mayo, 2005). Thus, severity assessments are always carried out post-data with respect to a specific inference.

It should be noted that although the above is mainly about experiments and statistical tests, the notion of severity can be employed in discussing the error characteristics of research tools. In fMRI, the complex workings of neuroimaging tools and processes require scrutiny just as well as statistical models and analyses. The normalization procedure in fMRI to correct for neuroanatomical variability is just one of these procedures which need to be analyzed with respect to their error characteristics. One needs to assess the error probabilities, stemming from neuroanatomical variability, associated with inferences of the form "participants had significant activity in brain region X as they performed cognitive task C". That is, we have to do a good job at probing the ways in which we may have been wrong in inferring that activity was in region X. Several participants may have had activity in region Y and we need to take into account this possible source of error. If we want to have severe tests of hypotheses in fMRI experiments, we need to assess the probabilities associated with this and other kinds of errors. If these error probabilities are found to be low, then we can statistically argue that it was very improbable that we committed those errors. Thus, by utilizing these error probabilities, we can improve the reliability of our inferences.

In order to control for and minimize errors due to neuroanatomical variability, some scientists suggest using probabilistic spaces based upon combining data from hundreds of neuroanatomical scans. One probabilistic space used in normalization is the Montreal Neurological Institute (MNI) template based on hundreds of brain images (Mazziotta et al., 1995). This a constructive step toward approximating more closely the actual error probabilities associated with inferences to hypotheses about relationships between cognitive performance and activation in certain brain regions. Of course, there may be biases in this atlas and there is always room for improvement. Indeed, other groups of researchers have been working in collaboration with the Mazziotta group for updates. Duncan (2009) in the Discover magazine reported that the team of researchers studied

scans of 450 brains and used hundreds of thousands of images taken from 7,000 people around the world as they updated the atlas. A more recent probabilistic brain atlas was developed by a group of researchers at the University of California, Los Angeles (Shattuck et al., 2007). In this probabilistic atlas, 56 brain structures were labeled in anatomical MRI scans and, for every voxel, probabilities of belonging to each of those 56 structures were calculated.

In order to illustrate how the error-statistical reasoning would work on the basis of a probabilistic brain atlas, let us assume that we do an fMRI experiment on the neural substrates of working memory. Previous studies have shown that the caudate nucleus, a brain structure that is connected with the thalamus and higher cortical structures, is involved in certain working memory tasks (among others, see Baier et al., 2010; Provost et al., 2010). We test the experimental hypothesis H, "the caudate nucleus is involved in the performance of working memory task W." As above, this experimental hypothesis would be linked to the alternative hypothesis H_a: $\mu_1 - \mu_0 > 0$ to be tested against the null hypothesis H_0: $\mu_1 - \mu_0 = 0$ where μ_0 designates mean activation in the caudate nucleus in the control condition and μ_1 designates mean activation in the caudate nucleus in the experimental condition in which participants perform working memory task W. Now, the conditional in the *modus tollens*, used above to illustrate Duhem's problem, will be: If H: the caudate nucleus is involved in the performance of working memory task W, then *e*: as participants perform W, there will be a significantly higher level of activity in the caudate nucleus compared to the control condition where they do nothing or perform a simpler or different task. Roughly speaking, there are two possible results of the experiment; one result is obtaining not-*e*, i.e. no significant activation in the caudate nucleus as participants perform W. The other possibility is that we do obtain significant activation in the caudate nucleus, which we can denote simply as *e*.

Let us first discuss the case of obtaining not-*e*; that is, we carry out the experiment and we observe no significantly higher activation in the caudate nucleus as participants perform working memory task W. So, we do not reject the null hypothesis and, following the *modus tollens*, we conclude that the experimental hypothesis H is not true, i.e. the caudate nucleus is not involved in the performance of working memory task W. Against

this conclusion, one can rightly raise the Duhemian objection and suggest that the inferential procedure should rather be like this: If H_1, H_2, H_3,…, H_n and A_1, A_2, A_3,…,A_n, then *e*. Not-*e*. Therefore, not-H_1 or not-H_2 … or not- H_n, or not-A_1 or not-A_2… or not- A_n where H_1 through H_n and A_1 through A_n are auxiliary hypotheses and assumptions involved in our experiment that yielded not-*e*. H_1 would be the experimental hypothesis of interest; namely "the caudate nucleus is involved in the performance of working memory task W" and *e* would be the prediction "as participants perform W, there will be significant activation in the caudate nucleus." This time around, the assumption, A_1, is replaced by the normalization procedure *based on a probabilistic brain template*, so we now have an empirical measure of the neuroanatomical variability across participants. This means that we have at least the approximate error probabilities associated with mappings between patterns of neural activity and neuroanatomical regions. If these error probabilities are low, then when not-*e* is obtained, that is, no significant activity in the caudate is observed, we can rule out the normalization procedure as a possible reason for obtaining not-*e*. Of course, there may be other reasons for obtaining not-*e*, but other components of neuroimaging experiments can, and should, be similarly scrutinized with respect to their error probabilities or error characteristics. This follows Mayo's solution of Duhem's problem: "Before experimental results can speak for or against a hypothesis under test, it is necessary to check and estimate the extent of any errors along the way—regarding the data and the auxiliaries" (1997; p. 231). If the normalization procedure incorporates empirical assessments of neuroanatomical variability, then we have at least some idea on the error probabilities associated with inferring that significant activity is in a certain brain region. If these error probabilities are low, then we can rule out the possibility of blaming the normalization procedure for obtaining not-*e*.

Let us now look at the second possibility; we carry out the fMRI experiment and the results show that when participants perform W, there is a significantly higher level of activation in a group of voxels that we identify as the caudate nucleus, so we get *e*. Therefore, assuming that the experiment was carried out without any serious flaws, we reject the null hypothesis H_0: $\mu_1 - \mu_0 = 0$ and accept the alternative hypothesis H_a: $\mu_1 - \mu_0 > 0$ so we conclude that the caudate nucleus is in fact involved in the performance of working

memory task W. Of course, this result, namely experimental data that agree with our hypothesis, by itself will not be sufficient to infer that our experimental hypothesis is true. Among the ways in which one could object to this conclusion is saying that the observed result does not necessarily mean that there really was activation in the caudate nucleus. This is because it can be argued that some participants' brains may have been different enough anatomically that although the results show activation in the caudate nucleus as identified by the Talairach space, it is possible that several participants had activation in the internal capsule, a brain structure adjacent to the caudate nucleus. Such an objection can be addressed if we have used a probabilistic brain atlas. That is, as we analyze our data we can take into consideration the error stemming from neuroanatomical variability. The probabilistic brain atlas would give us the probability of correctly identifying a group of voxels as a specific structure on the basis of hundreds of anatomical brain scans. In the case at hand, we are interested in significant levels of activity in the caudate nucleus, so we can consult the probabilistic atlas about the group of voxels we found to be activated. The atlas tells us how often that group of voxels has been identified as the caudate nucleus; specifically, 92% of the time it has been identified as the caudate nucleus, 6% of the time as internal capsule, 1% of the time as anterior horn of lateral ventricle, and less than 1% as other regions (Mazziotta et al., 1995). This information helps assess how often we may misidentify these brain regions. Assuming that no functional activation would be detected by fMRI in the lateral ventricle, we can say that the probability of misidentifying the group of voxels in this experiment as the caudate nucleus was 7% and this probability comes from hundreds of anatomical brain scans. Very rarely do fMRI experiments have more than 15 or 20 participants, so the probability of misidentification may be even lower in our experiment. The reason is that larger numbers of participants in an experiment would probably increase the chances of neuroanatomical variability producing errors in identification of brain regions. In the worst case, the probability of misidentifying the group of voxels was 7%, which is not too high, so with the error probability associated with this type of error at hand, we can say that it was improbable that this specific error was committed in the given experiment. Therefore, we can statistically rule out neuroanatomical variability as a serious source of error and infer more reliably and

accurately where activity really took place in the brain.

The example of neuroanatomical variability across participants is only one among many aspects of a neuroimaging experiment that need to be scrutinized in this kind of error-statistical analysis. In order to deal with all the errors or flaws that preprocessing techniques may introduce, they should be analyzed for their error probabilities or error characteristics. Let me illustrate how we can begin carrying out such error-statistical analyses by looking at a specific preprocessing technique. Spatial filtering, or smoothing, is a computational procedure applied to raw fMRI data in order to reduce the noise due to non-task related sources of variability such as heart rate or respiration. If successful, one effect of smoothing is that noise is averaged out while the task-related signal is left unaffected (Lazar, 2008; p.48). Essentially, smoothing combines and spreads the data observed in multiple voxels, which ends up "blurring" the neuroimages. The fMRI signal as measured across voxels exhibit spatial correlations, that is, if a voxel is active, then with high probability nearby voxels are also active. There are many things that may be the reason for this; one probable reason is that adjacent regions of the brain may also be functionally similar. In addition, brain regions are highly connected with nearby regions, so when one group of voxels is active, this may cause nearby voxels to also be active. Using a spatial filter, which corresponds to spatial correlation expected to occur because of functional similarity and connections of brain regions, greatly improves the functional signal-to-noise ratio (SNR) (Huettel et al., 2004; p.277). A common blurring technique is applying a Gaussian filter, which can be characterized by the measure called "full width at half maximum (FWHM)" of the observed signal, which is defined as $2\sqrt{2\log}\sigma$ for a Gaussian distribution that has variance σ^2 (Lazar, 2008; p.48). When a Gaussian filter is applied to fMRI data, it spreads the observed signal over other voxels that are nearby. Spatial filters may be wide or narrow; narrow filters combine data from a few voxels, whereas wide filters combine data across many voxels (Huettel et al., 2004; p.276). The width of the spatial filter applied in an experiment is expressed in millimeters at half the maximum value of the fMRI signal. For example, a filter width of 10 mm FWHM combines data from approximately 3 voxels (Lazar, 2008; p.48). As the width of the spatial filter increases, more smoothing is applied combining data from more voxels.

Lazar (2008) cites two benefits of applying spatial filters. One benefit is that spatial filtering improves the functional SNR, thus making the fMRI experiment more powerful in detecting task-related signals. The other benefit of spatial filtering is that it makes the data have a distribution closer to a normal distribution. Thus, spatial filtering is supposed to improve the quality of data for statistical analyses. Huettel and colleagues (2004) suggest that spatial filtering improves the validity of statistical analyses. They point to the fact that in a volume of data, there may be as many as 102,400 voxels and if the threshold for significance is set at .05, then, assuming independence of voxels, when we carry out significance tests for each voxel to determine whether or not it is active, as many as 5,000 voxels may be detected as active due to mere chance. They write that if spatial filtering is applied, then "there may be many fewer local maxima that exhibit significant activity" (ibid., p. 277). Thus, spatial filtering is helpful also for the multiple testing problem and allows researchers to use correction techniques less conservative than the common, highly conservative technique of Bonferroni correction.

However, as Lazar (2008) describes, spatial filtering also has certain disadvantages. Researchers have to be very careful in choosing the width of the spatial filter they will employ. If the width is not appropriate, the filter applied may have negative effects on statistical analyses of preprocessed data. If the filter employed is too wide, that is, data from many voxels are combined, then data from regions that are not active may be included. This may occur when there is significant activation in a very small brain region, but if data from nearby nonactive voxels are combined in the filter, then the activation in the small region may be smoothed out and rendered undetectable. On the other hand, if the applied filter is too small, it will not be effective in improving the SNR, so nothing would be gained while spatial resolution would be degraded. Another disadvantage of spatial filtering is that it may cause the merging together of brain regions that are functionally different (ibid.). This may lead to contradictory fMRI findings in different experiments or even in different kinds of analyses of the same data set. All these disadvantages of spatial filters may introduce errors that may influence the experimental findings independently of the truth or falsity of the experimental hypotheses that fMRI experiments are meant to test. Thus, it is possible for researchers to obtain results that agree with their experimental

hypothesis not because the hypothesis is true, but because they applied a spatial filter to their data. Fransson et al. (2002) demonstrated clearly how this can happen.

In an experiment, Fransson and his colleagues (2002) asked participants to do an episodic memory encoding task as fMRI data were collected. Then, they applied two different types of analyses to the same data set; one type of analysis included no spatial filtering, and the other included a spatial filter with a width of 8 mm, a filter size commonly used in fMRI research. The analysis with spatial filtering yielded significantly high amounts of activation in the hippocampus, whereas the analysis without spatial filtering did not. This result demonstrates that a finding may be obtained not necessarily because the hypothesis being tested is true, but rather because of some procedure applied to the data set in the preprocessing stages. As Franssen and his colleagues state, "the results of an fMRI study appear to be crucially dependent on the approach chosen for post-acquisition data processing and analysis" (ibid., p. 981)., Lazar (2008) notes that, because of these and similar disadvantages, some research groups do not include any spatial filtering as part of their preprocessing protocols. This may be too radical a choice, because not applying any spatial filters may lead to an experiment with low power where task-related signals of interest may go undetected. As with any other aspect of a given fMRI experiment, trying to find out the error characteristics of spatial filtering is a more beneficial approach than blindly applying spatial filters or not applying any spatial filter at all. Lazar advocates an approach in a similar vein. She suggests analyzing data sets without spatial filtering and then analyzing the same data sets several times with spatial filtering of varying widths. The results of these analyses can show us how dependent the experimental results are on spatial filtering. This could also tell us how often the inferences we draw from data are influenced by procedures like spatial filtering. Indeed, analyses of this type must be expanded to include other preprocessing procedures, such as distortion correction, temporal filtering, etc. Ideally, each preprocessing procedure would be analyzed with respect to its error characteristics and how it may influence the outcomes of statistical analyses.

One paradigm for the assessment of the effects of preprocessing procedures has been proposed by Strother and his colleagues. The paradigm is called the "nonparametric

prediction, activation, influence, and reproducibility resampling" or NPAIRS framework (Strother et al., 2002; LaConte et al., 2003). This paradigm makes use of the notion of cross-validation where an fMRI data set is split in 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 data, which is designated as the "test" data. This process is repeated in a second run but with the training and test data switched, that is, in the second application of the process, test data are used for training and training data are used for testing. Thus, researchers can assess the prediction accuracy of their models. Reproducibility of the findings is assessed by comparing the results of statistical analyses on both halves of the data across several runs. The flexible nature of this analysis paradigm allows researchers to assess the effects of different types of preprocessing protocols on fMRI data. For example, LaConte et al. (2003) compared the effects of different preprocessing protocols on prediction accuracy and reproducibility. Across several runs of the split half process described above, they applied different preprocessing protocols, which they called analysis chains. Each analysis chain included different levels of preprocessing of the data; one chain included no preprocessing procedures, whereas others included normalization and different degrees of spatial filtering, e.g. one chain applied a narrow filter and another chain applied a wide filter. Then, they did final statistical analyses on data sets that came from these different analysis chains in order to assess the effects and contribution of different preprocessing protocols, or analysis chains, on prediction accuracy and reproducibility. The results showed that the greatest improvement in improving prediction accuracy and reproducibility came with spatial smoothing.

However, as LaConte and his colleagues (2003) note, there are no general pre-data guidelines for what the optimal preprocessing protocol would be for all experiments. One reason for this is that 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 experiment, etc. Therefore, the evaluation of preprocessing protocols with respect to their error characteristics and/or their effectiveness will have to be done on a case by case basis.

This is very much in line with the piece-meal approach of the error-statistical account as well as the essential notion that the severity of a test is always assessed post-data in terms of a specific hypothesis, the data set at hand, and the experiment that generated that data set. When we place the severity function, SEV(T, $\mathbf{x}_0$, H) in the context of fMRI experiments, we can think of the preprocessing protocol as another aspect of T, i.e. the experiment that generated the data. As Lazar (2008) and LaConte et al. (2003) suggest, we can apply different preprocessing protocols to the same set of raw fMRI data and then do statistical analyses on the data sets yielded by the different preprocessing protocols. In this way, we can assess the effects these protocols have on the results of tests on the same data set. The results of these analyses can be helpful in finding out the error probabilities associated with different preprocessing protocols and, in turn, how these affect the severity of the whole experiment as a test of the experimental hypothesis of interest. One crucial point is that as practitioners become more aware of the errors that preprocessing procedures may introduce, they start devising methods of identifying and controlling for the ways in which these errors arise in fMRI experiments. The NPAIRS framework is a good example of this kind of work. The error-statistical notions of error probabilities and severe tests can aid this methodological trend by supplying useful conceptual machinery and additional criteria for the assessment of errors and inferences in fMRI studies.

In the error-statistical framework, we can break down a neuroimaging study in piece-meal fashion into its component parts and procedures from experimental design to initial data collection, and from preprocessing to statistical modeling and hypothesis tests. We can then assess the error probabilities, or error characteristics, associated with each component, as was done above regarding the use of probabilistic brain atlases or different types of preprocessing protocols. Thus, on a case by case basis, we can assess how these procedures may introduce errors and to what extent, if at all, they may influence the results independently of the truth or falsity of experimental hypotheses of interest. The component parts and procedures of an fMRI experiment can also be thought of as factors that determine the severity of that experiment as a test of the specific hypothesis of interest. If the components of an experiment may introduce errors with high probability or if they have characteristics prone to create biases in data, then this renders the experiment a low

severity test of the specific hypothesis of interest. The use of statistical models and significance tests is one of the most important components of an experiment. As described earlier, most fMRI experiments are done to learn about the truth or falsity of experimental hypotheses about relationships between neural activity and performance of cognitive tasks. These hypotheses are tested in significance tests which are formulated in the context of a statistical model of fMRI data. If a researcher wants an experiment with low error probabilities that can put the hypothesis of interest to a severe test, then one thing, among other things, that she has to make sure is that the hypothesis tests are carried out without flaws. To achieve this, the researcher must have adequate statistical models of fMRI data. If we recall the above Duhemian formulation with the auxiliary hypotheses and assumptions, we can think of the assumption that the statistical model chosen for an fMRI data set is adequate as one of those auxiliary assumptions. When an fMRI result does not fit the hypothesis of interest, Duhemian objections, similar to the ones above about neuroanatomical variability or inadequate preprocessing protocols, can be raised by saying that this result was obtained because the data were not adequately modeled. The crucial point here is this; the auxiliary assumption that researchers have modeled the data adequately can be tested for in the error-statistical approach. Here, I briefly discuss how the error-statistical approach can aid in statistical modeling of fMRI data.

The general linear model (GLM) is commonly used in fMRI research (Huettel et al., 2004; Lazar, 2008) and the factors in the GLM represent the hypothesized components of the data. Given the experimental data and model factors, researchers calculate the combination of factor weights that minimize the error term. If there is only one model factor, then the GLM is identical to a correlation analysis; if there is only one model factor with two levels, then the GLM is identical to a t-test. The form of the GLM can be expressed in the equation: $Y = X\beta + \varepsilon$ where Y is the preprocessed fMRI data, which may be represented in a matrix of the time series data from all voxels, so it will have one column for each voxel and one row for each time point (Lazar, 2008; p.83). X represents the model factors and can be expressed in terms of a design matrix representing the stimuli or tasks presented to the participants during the course of the experiment. For example, pictures that were shown to participants, tasks they were asked to perform, and the time

points at which these were presented would be included in the design matrix. β represents the unknown coefficients of the model factors and ε represents the error, which is assumed to be normally distributed with mean zero and variance σ^2 (ibid.). The GLM in this form is a basic example of how statistical tests are thought of in the fMRI literature. Statistical tests are conceived as tools to find out which experimental manipulations, i.e. factors, have the greatest effect on the preprocessed fMRI data. In other words, statistical tests are designed to discover whether or not manipulations of cognitive tasks produce significant increases in activation in the brain as a whole, or certain regions of the brain.

The GLM, same as any other statistical model, comes with a set of probabilistic assumptions about the data generating mechanism. These are: 1) The data Y is normally distributed; 2) The process that generated the data Y is an independent process; 3) The expectation of data Y is linear in X; 4) The variance of data Y is homoscedastic, i.e. variance of Y is free of factors X (Spanos, 1998). Functional MRI researchers use the GLM to model data where they assume that: 1) Raw fMRI data can be modeled as the sum of separate factors and additive Gaussian noise, 2) Each factor may vary independently across voxels, and 3) Gaussian noise is independently and identically distributed (Huettel et al., 2004; p.342). Of course, the verification of these assumptions is of crucial importance in order to establish the validity of statistical inferences. However, Lazar (2008) states that the assumptions of the GLM in the context of fMRI "are surely unrealistic and hence violated in practice…" (p.85).

Lazar (2008) discusses two general questions that arise in model validation in fMRI; one is a question about which model, among many alternative models, should be chosen to fit to the brain as a whole. Lazar states that the difficulty here is that the notion of fit does not have a precise definition in this context. The other question is about whether or not the same model should be fit to every voxel in the brain. Given the variability of fMRI data across voxels, it seems that if the same model is fit to every voxel, some voxels will be underfit while others will be overfit. On the other hand, if different models are fit to different voxels, some necessary statistical procedures cannot be used. For example, detecting contiguous groups of active voxels is crucial for any experiment; one way in

which this can be achieved is by applying random field thresholding. However, this thresholding technique cannot be used if different models are used for different voxels.

Even though the above problems are indeed serious, the fundamental problem in modeling fMRI data is the fact that assumptions of standard models such as GLM are violated in the practice of fMRI research. For example, the independence assumption of GLM is violated in fMRI experiments. One reason for this may be that regions of the brain are densely connected with each other and when one region is activated this causes activations in nearby regions as well. Thus, the process that generates fMRI data is not always an independent process and this may threaten the validity and reliability of statistical inferences. This difficulty may be one of the factors responsible for the relatively high incidence of contradictory findings in the fMRI literature. Lazar (2008) calls attention to several drawbacks in fMRI analyses that are caused by problems of model validation. Some of these drawbacks are misspecification of models, choosing oversimplistic models due to a lack of criteria for systematic evaluation of models, and improper choice of models on the basis of number of active voxels, where a model is considered to be a better model if it detects more active voxels. All these may introduce biases or flaws in the data analyses and significance tests, increase error probabilities, and lead to erroneous inferences.

In this environment where fundamental problems of data modeling threaten the validity and reliability of inferences, as recognized by fMRI researchers like Petersson and his colleagues (1999) and Lazar (2008) as well as others, the error-statistical approach to model validation proposed by Mayo and Spanos (2004; 2010; 2011) can be useful. A central aspect of this approach is misspecification (M-S) testing, which includes methods of testing the model assumptions about the data generating mechanism. Another essential element of M-S testing is respecification; if assumptions of a model are violated, iterative procedures are applied to accommodate flawed assumptions in respecified models. In the end, a statistically adequate model of the data at hand is obtained, which can support reliable inferences about the hypotheses of interest. Another advantage of M-S testing is that it distinguishes between problems of model specification and problems of model selection where researchers select a model from an assumed family of models. M-S testing

provides a method for developing statistically adequate models of given data sets. Once we have a data set, say preprocessed fMRI data, we can proceed by what Mayo and Spanos (2004) call the probabilistic reduction approach in which we think of the set of all possible statistical models of the mechanism that generated the data. Every statistical model is a set of probabilistic assumptions about the data generating mechanism and these assumptions can be grouped under three broad categories: distribution, dependence, and heterogeneity. Given a specific fMRI data set, we can start the specification process by asking general questions about the data set, such as 'are the data independent over time?', 'are the data from different voxels, or different regions of interest, independent?', 'what is the distribution of the data? e.g. normal or skewed?', 'are the data from different voxels, or different regions of interest, identically distributed?' The answers to these questions will eliminate certain possibilities for the model to be chosen. For example, as has been noted before, in fMRI, data from neighboring voxels are not independent. In fact, often there is spatial correlation between data from adjacent voxels as is to be expected given the highly connected anatomy and functioning of the brain. Thus, any model that cannot accommodate this dependence in the data would be eliminated as a potential model. Obviously, given the large size and complexity of fMRI data sets, the application of M-S testing to fMRI data will be a serious undertaking. However, when researchers proceed according to the probabilistic reduction approach, they can develop statistically adequate models of data, which would control and minimize error probabilities associated with modeling and significance tests in fMRI. Thus, when M-S testing is properly applied and adequate models of data are specified, no Duhemian objections can be raised about issues of modeling. This would mean that another component of the fMRI experiment, namely statistical modeling of data, is ruled out a source of error in the experiment as a whole.

In this paper, I have demonstrated how the error-statistical approach can help us better understand, formulate, and tackle Duhemian problems in fMRI. Error-statistical analyses provide estimates of the probability of making erroneous inferences due to problems in different stages of an experiment. By looking at these probabilities for any given experiment, we can more accurately assess the reliability of our inferences. The error-statistical approach can give us the kind of characterization necessary for complete

and accurate assessments of inferential steps and we can go the inferential distance, as it were. In other words, with philosophical and statistical arguments using the notions of error probabilities and severe tests, we can counter those who are skeptical of the epistemic value of fMRI findings. Duhem's problem provides the most useful conceptual framework in which we can describe general methodological and inferential problems in functional neuroimaging. The error-statistical account helps us clearly formulate and tackle these problems. Thus, we can determine the kind of knowledge functional neuroimaging can reliably provide and the conditions under which it can provide it without prematurely conceding to skepticism or pessimism about the epistemic value of neuroimaging findings.

References:

Baier, B., Karnath, H.O., Dieterich, M., Birklein, F., Heinze, C., Muller, N.G. (2010). "Keeping Memory Clear and Stable—The Contribution of Human Basal Ganglia And Prefrontal Cortex To Working Memory." *The Journal of Neuroscience*, *30*, 9788 –9792.

Bogen, J. (2010) "Theory and Observation in Science." *The Stanford Encyclopedia of Philosophy (Spring 2010 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2010/entries/science-theory-observation/>.

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

Duncan, D. E. (May, 2009). "Looking at Stress—and God—in the Human Brain." *Discover Magazine.*

Fransson, P., Merboldt, K.D., Petersson, K.M., Ingvar, M., Frahm, J. (2002). "On the Effects of Spatial Filtering—A Comparative fMRI Study of Episodic Memory Encoding at High and Low Resolution." *NeuroImage*, *16*, 977–984.

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.

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.

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

Mayo, D. (1997). "Duhem's Problem, the Bayesian Way, and Error Statistics, or 'What's Belief Got to Do with It?'" *Philosophy of Science, 64* (2), 222-244

Mayo, D. (2005). "Evidence as Passing Severe Tests: Highly Probable versus Highly Probed Hypotheses" in Achinstein, P. (ed.) *Scientific Evidence: Philosophical Theories & Applications.* Baltimore, MD: The Johns Hopkins University Press.

Mayo, D. & Spanos, A. (2004). "Methodology In Practice: Statistical Misspecification Testing." *Philosophy of Science, 71*, 1007-1025.

Mayo, D. & Spanos, A. (2010). *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity of Science*. New York, NY: Cambridge University 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.

Mazziotta, J. C., Toga, A. W., Evans, A., Fox, P., & Lancaster, J. (The International Consortium of Brain Mapping, ICBM). (1995). "A Probabilistic Atlas of the Human Brain: Theory and Rationale for Its Development." *Neuroimage, 2*, 89-101.

Petersson, K.M., Nichols, T.E., Poline, J.B., & Holmes, A.P. (1999). "Statistical Limitations in Functional Neuroimaging I. Non-inferential Methods and Statistical Models." *Philosophical Transactions of the Royal Society of London, 354*, 1239-1260.

Provost, J.S., Petrides, M., & Monchi, O. (2010). "Dissociating The Role Of The Caudate Nucleus And Dorsolateral Prefrontal Cortex In The Monitoring Of Events Within Human Working Memory." *European Journal of Neuroscience, 32*, 873-880.

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

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

Shattuck, D.W., Mirza, M., Adisetiyo, V., Hojatkashani, C., Salamon, G., Narr, K.L., Poldrack, R.A., Bilder, R.M., Toga, A.W. (2007). "Construction of a 3D Probabilistic Atlas of Human Cortical Structures." *NeuroImage, 39 (3),* 1064 – 1080.

Spanos, A. (1998). *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*. Cambridge, UK: Cambridge University Press.

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.

Talairach, J., & Tournoux, P. (1988). *Co-planar Stereotaxic Atlas of the Human Brain*. Thieme Medical Publishers, New York.