Brit. J. Phil. Sci. **61** (2010), 265–278

Images Are Not the Evidence in Neuroimaging

Colin Klein

ABSTRACT

fMRI promises to uncover the functional structure of the brain. I argue, however, that pictures of 'brain activity' associated with fMRI experiments are poor evidence for functional claims. These neuroimages present the results of null hypothesis significance tests performed on fMRI data. Significance tests alone cannot provide evidence about the functional structure of causally dense systems, including the brain. Instead, neuroimages should be seen as indicating regions where further data analysis is warranted. This additional analysis rarely involves simple significance testing, and so justified skepticism about neuroimages does not provide reason for skepticism about fMRI more generally.

1 Introduction

Neuroimages—colorized pictures of 'brain activity'—are the most well known products of fMRI experiments.[1] They are often taken to be evidence for *functional hypotheses*: that is, evidence that a given brain region plays a particular causal role during the performance of a cognitive task.[2]

[1] In this paper, 'imaging' and 'fMRI' will refer to BOLD (blood-oxygen-level dependent—see Section 2) fMRI, and 'neuroimages' to the products of fMRI described in Section 2. Much of what I will say will carry over to other imaging modalities, but note that the argument does depend on detailed considerations about the generation of the BOLD response.

[2] More precisely, by 'functional hypothesis' I will mean claims of the form 'The function of A is to F in order to E', where A is some region of the brain, F some activity that it performs, and E

doi:10.1093/bjps/axp035

Advance Access published on August 7, 2009

I will argue that neither neuroimages nor what they depict provides evidence for functional hypotheses. Further, I will argue that skepticism about neuroimages can be grounded in well-known problems with the use of null hypothesis significance testing (NHST). The problems with neuroimages are thus conceptual, rather than merely practical, and cannot be easily avoided. In this sense, I am adding to a long-established skeptical tradition in the philosophical literature on neuroimaging.[3]

Yet this does not mean that we should be skeptical about *neuroimaging*—that is, about fMRI and the associated techniques. The overwhelming majority of contemporary fMRI experiments present more evidence than is presented in neuroimages. This evidence rarely consists of simple NHSTs. As such, this further evidence is not touched by skepticism about neuroimages. In most cases, fMRI provides precisely the sort of evidence that opponents of NHSTs would urge us to seek.

I conclude that we should view neuroimages as auxiliaries to evidence, rather than evidence proper. Neuroimages indicate brain regions in which further analysis may provide warranted and fruitful evidence for functional hypotheses. It is this further analysis that provides the evidence, rather than the neuroimages themselves. Neuroimaging may thus remain a fruitful technique even if the status often attributed to neuroimages is unjustified.

2 Neuroimages Are Statistical Maps

Differences in brain activity when subjects perform different cognitive tasks might be thought, all things being equal, to provide evidence for the functional role of brain areas. It is this insight that drives much contemporary neuroimaging, and many take neuroimages to provide evidence for functionally relevant brain activity.

fMRI works by tracking the changes in blood oxygenation that occur after increased local brain activity. These changes in local oxygenation can be detected by a properly sequenced MRI scanner, and provide an indirect measure of increased neural activity.[4] Changes in this blood-oxygen-level dependent (BOLD) MR signal are the primary data produced by fMRI.

some overall activity toward which the *F*-ing of *A* contributes. (For example, the proposition that *the function of V5/MT is to detect moving objects during normal vision* is a functional hypothesis.) In this sense, I will be most concerned with what are sometimes called 'causal' or 'Cummins' functions (after Cummins [1999]). I will assume that true functional claims imply that if *A* had not *F*-ed on a particular occasion, then *E* would either have failed to happen or happened in some different manner. Functions are thus able to enter into explanations of why agents have particular psychological capacities.

[3] See, for example, (Uttal [2001]; Hardcastle and Stewart [2002]; Coltheart [2006]).

[4] Increased brain activity requires an increase in oxidative metabolism, and so it causes changes in local blood oxygenation. These changes have characteristic effects on a magnetic resonance signal. Deoxyhemoglobin is paramagnetic and so it causes local spin dephasing in transversely magnetized hydrogen molecules. This dephasing results in a decrease in the net MR signal

That fMRI is an indirect measure is in itself unremarkable, and should not engender skepticism. Neuroimages are not simple pictures of BOLD signal differences, however. Quantitative signal magnitudes are effectively uninterpretable on their own, as there is no general mapping from BOLD signal to functional significance of neural activity.[5] Further, the BOLD differences associated with brain activity are small, noisy, and temporally complex. In lieu of quantitative information, neuroimages instead show maps of regions where there was a *statistically significant* difference in BOLD signal between task conditions.

To produce such maps, the BOLD signal in each subregion is subjected to a NHST between conditions of interest. NHSTs consist of two steps. First, one computes the likelihood that one would observe a given set of data if the *null hypothesis* were true. The null hypothesis is the proposition that an experimental condition had no real effect on the observed MR signal, and so that the neural activity in a region remained unchanged while the subject performed different cognitive tasks. This value, representing the likelihood of observing data conditional on the null hypothesis, is referred to as the p-value. In the second step, one compares the p-value to a predetermined significance level α. If the p-value is lower than α, the data is *statistically significant*. A significant result is one that would be unlikely to be observed if the null hypothesis were true—with an α of 0.01, for example, one could expect to see significant data in only about 1 of 100 observations in regions where the null hypothesis was true.

Neuroimages are produced by performing NHSTs in each three-dimensional subregion (or *voxel*) of the data.[6] The results are plotted as *statistical parametric maps* (SPMs). SPMs show those voxels in which the p-value for that region was significant, i.e., less than or equal to the significance level α. Typically, a range

from a region, allowing an MRI scan to detect local differences in blood oxygenation. I omit considerable detail; Buxton ([2002]) provides a useful introduction to MRI technology that covers both structural and functional MRI.

[5] BOLD responses vary in strength between functional regions (Nair [2005], p. 234). Signal strength varies with nonfunctional parameters (like magnet field strength). It's not clear what the mapping between strength and relevant neural parameters should be, and it is unlikely that there is a single such mapping (Nair [2005]). Logothetis ([2008]) details one important source of interpretive difficulty, the complicated relationship between the BOLD signal and the mass action of excitatory–inhibitory networks. The difficulties involved in interpreting quantitative magnitudes are one reason why contemporary neuroimages are presented in the way that they are (Buxton [2002], p. 423).

[6] In the simplest case, this is done via a t-test of the average magnitude following each task condition. The t-tests do not take into account facts about the shape of the hemodynamic response, however, and they require block designs that do not allow for rapidly interleaved task conditions. In most experiments, therefore, the signal from each voxel for each task is analyzed via a general linear model, and the signal from each voxel is fitted to a canonical model of the hemodynamic response convolved with a step function representing task epochs. This model includes at least one free parameter for the amplitude of the response. More complex models include additional free parameters for the time of onset and dispersion of the hemodynamic response function (see, e.g., Chapter 4 of Sarty [2007]).

of colors is used to represent the magnitude of the p-value calculated for a region, with brighter colors indicating lower p-values. SPMs thus summarize the results of thousands of simultaneous significance tests, showing the areas in which our data permit us to reject the null hypothesis of no difference in activation between conditions.

Neuroimages are SPMs overlaid on anatomical images of subjects' brains. Neuroimages are not maps of activation per se, but rather maps of places where we may be confident that the resemblance between data and a stereotyped pattern of activation is unlikely to be the result of chance fluctuations from a true zero signal. So, neuroimages do not show differential activity. They show places where (*ceteris paribus*) the data warrant confident assertion of a pattern of differential activity.

Many people, especially nonspecialists, take neuroimages to be especially good evidence for functional claims (Dumit [2004]; McCabe and Castel [2008]). Working scientists are typically more cautious. Nevertheless, I argue that they usually take neuroimages (and what they represent) to be at least weak evidence for functional claims (see Mole and Klein [forthcoming] for a defense and discussion of this claim). I argue that this is mistaken: neuroimages do not provide even weak support for functional hypotheses.

3 The Skeptical Argument

3.1 Evidence and neuroimages

fMRI evidence results from a chancy sampling of the world, and requires a probabilistic analysis. I will assume that an updating of odds on a functional hypothesis H_a relative to a null hypothesis of functional unimportance H_0 given some evidence D is rational just in case

$$\frac{p(H_a \mid D)}{p(H_0 \mid D)} = \frac{p(D \mid H_a)}{p(D \mid H_0)} \times \frac{p(H_a)}{p(H_0)}.$$

The *likelihood ratio* $p(D \mid H_a)/p(D \mid H_0)$ gives a measure of the degree to which D supports H_a over H_0. A likelihood ratio greater than 1 indicates confirmation, while a ratio less than 1 indicates infirmation. Whether neuroimages are appropriate for confirming functional hypothesis thus requires consideration of three factors: the nature of the evidence D, and facts about the conditional probabilities $p(D \mid H_a)$ and $p(D \mid H_0)$. We will rarely be able to put precise numbers on the latter probabilities, but we can say useful things about the rough relationship between them.

Skepticism about neuroimages amounts to the proposition that the likelihood ratio of a functional hypothesis to its null is always very low when we treat an SPM as data. The precise form of this skepticism depends on which of the three ways of construing D we choose. First, D could be the fact that there

was *increased activity*: that is, the fact that there was more brain activity in one condition than in another. Second, D could be the fact that there was a *statistically significant difference* in activity: not just difference, but difference that was statistically detectable. Third, D could be associated with the *actual time-course of some statistically significant data*: not just the fact of significance, in other words, but the fact that the difference was significant *and* took thus-and-such shape. Each of the three ways of reading D makes neuroimages problematic as evidence.

3.2 The problem of causal density

Suppose D is the fact that there was task-related differential brain activity. The problem: there is decent reason to believe that any task will have widespread effects on the brain. These effects will be small and functionally insignificant—but nevertheless, they will be present. Which means that both $p(D \mid H_a)$ and $p(D \mid H_0)$ are high in each area of the brain, and the likelihood ratio is close to 1. Given this, D is uninformative.

fMRI is relatively insensitive: everyone agrees that there are real differences in brain activity that get lost in noise. But suppose we were able to make our fMRI experiments arbitrarily sensitive, so that even small differences in brain activity became detectable: There is a good argument that, were we able to do so, we should expect to find differential activity across the entire brain for any task. This is because brains are *causally dense* systems: systems in which there is a causal path between changes in any explanatory variable and most other variables. As Savoy notes, the brain is a densely interconnected system, one in which

> . . .there are only about five synapses between any two neurones in the brain. It is reasonably likely that the activity in any one neurone (or collection of neurones, given the spatial resolution of our non-invasive imaging techniques) is going to influence almost any other neurone, albeit weakly. (Savoy [2001], p. 30)

This is not to say, of course, that these widespread differences in activity will be functionally important. The point is merely that they are likely to be there.[7] But if differences are likely to be widespread, then the observation of difference is uninformative.

This is not an abstract worry. There is good evidence that fMRI experiments that look more carefully find more activity. Studies looking at the effect of

[7] This problem may be more or less compelling depending on what aspect of neural activity is tracked by the BOLD signal. In particular, if BOLD tracks differences in *synaptic* activity rather than spiking rates (as suggested by Viswanathan and Freeman [2007]), then small widespread differences in activity might be more likely. This is controversial; see (Nir et al. [2008]; Viswanathan and Freeman [2008]) for recent discussion.

increased sensitivity confirm that improving the signal-to-noise ratio of fMRI dramatically increases the number and extent of activated regions at the same α level. This is apparent in studies that increase the number of subjects (Savoy [2001], p. 30; Thirion et al. [2007]), the number of trials within a study (Huettel and McCarthy [2001]), and the field strength of the main magnet (Huettel et al. [2004], p. 237).

Put another way: if $p(D \mid H_0)$ is high, then the subthreshold activation simply indicates a failure of our instrument to detect a signal.[8] The fact that an imaging experiment *now* differentiates between activated areas thus seems like a fluke of instrumentation. In this case, as Hardcastle and Stewart complain, 'brain imaging seems to support localist assumptions because we aren't very good at it yet.' (Hardcastle and Stewart [2002], p. S78)

3.3 The problem of arbitrary thresholds

Suppose D is the fact that there was a *statistically significant* difference in the data. Claims of statistical significance are always relative to the choice of α. But, the skeptic argues, there is no rationally justifiable choice for α. The argument again relies on the causal density of the brain. The theoretical justification for choosing an α level depends on the desirability of reducing false positives. The actual rate of false positives is the product of α and of the base rate of true null hypotheses.[9] If everything in the brain is weakly connected to everything else, then every task should be expected to result in *some* difference in neural activation.[10] This means that the null hypothesis H_0 is always strictly false. But if there are no true nulls, then it is trivially impossible to have a *false* positive, no matter which α you choose. So if brains are causally dense, then any α will do, and the choice of one is arbitrary.[11]

Huettel et al. further note that the test statistics for individual voxels often change in a graded manner as one moves from region to region (Huettel et al. [2004], p. 246). Thresholding at any α inevitably creates artificially sharp barriers between 'active' and 'inactive' regions. This is why variation in α can

[8] Even if we take into account the directionality of the signal, we get at best a likelihood ratio of 2, since absent any other information the chance that an activated area activates in a particular direction is 0.5 (Meehl [1967], p. 111).

[9] The probability α is often considered to be the *absolute* chance of false positives in an imaging experiment (Sarty [2007], p. 66; Huettel et al. [2004], p. 345). This is a mistake: if there are no true negatives, the false positive rate is 0 no matter what the α level.

[10] Of course, there are clearly true nulls even in brain imaging: there cannot be task-related differential hemodynamic responses in either skull or ventricles, and so the null hypothesis is always true in voxels that contain only those tissues.

[11] 'Arbitrary' just means that there is no rationally compelling reason to choose any particular threshold. Distinguish this from the less plausible *ad hominem* charge that researchers pick thresholds that best support their conclusion (Lloyd [2002], p. 244). That appears to be false: though there is no widespread consensus, there is a fair bit of agreement on the acceptable ways of choosing an α; for a review of the standard possibilities, see (Huettel et al. [2004], pp. 343–51).

result in such dramatic differences in extent of activation: any choice of α makes a sharp distinction among what is a typically continuous variation in the underlying p-values. This means that different α values result in maps with dramatically different extents of activation. A conservative threshold shows very small activated areas, and a liberal threshold much larger ones.

Complaints about arbitrary thresholding are common in critiques of functional imaging (Hardcastle and Stewart [2002]; Uttal [2001], pp. 167–9; Roskies [2007], p. 870). Uttal, for example, complains about thresholds that 'a conservative assignment could hide localized activity and a reckless one suggest unique localizations that are entirely artifactual' (Uttal [2001], p. 168).[12] If choice of threshold is really arbitrary, then $p(D \mid H_0)$ and $p(D \mid H_a)$ will be similarly arbitrary. This means that there is never any rationally compelling way to fix the likelihood ratio and so no way to settle disputes about how strongly the data confirm a hypothesis.

Threshold choice can have theoretically important consequences. Savoy provides a graphical illustration of the point with data collected from subjects looking at flickering checkerboards (Savoy [2001], p. 28). Different thresholds generate images that show different patterns of activation. With a relatively high threshold, the map appears to indicate focal activity in V5/MT, a visual area associated with motion processing. At lower thresholds, all early visual areas (along with other regions of extrastriate cortex) show supra-threshold levels of significant activity. The debate between distributed and modular models of face recognition in part hinges on what to do with small activations in regions outside of fusiform face area (FFA). As Haxby et al. ([2001]) note, there are subthreshold activations outside of FFA that nevertheless contain enough information to recover whether a subject was looking at a face or a house. So it is consistent with the data that even small subthreshold activations might play a functionally important role in facial recognition.

3.4 The problem of vague alternatives

Suppose D is the actual difference in observed BOLD signal in a voxel. This is perhaps the most promising interpretation. Assume for a moment that functionally important areas will show a hemodynamic response, and that the data from some area do show such a canonical response. Then, one might argue, $p(D \mid H_a)$ is well defined and reasonably high: it will be equal to the statistical

[12] Uttal also criticizes the now-common practice of presenting gradations of color corresponding to different magnitudes of test statistic. This was developed in part to compensate for abrupt thresholding (Jernigan et al. [2003]). There is always a cutoff between active and nonactive voxels, however, so the threshold problem is not itself avoided by effect maps. Further, graded colorations can be misleading: it is easy to mistake them for a measure of strength of effect, rather than of strength of confidence that there was an effect.

power of the experiment. The likelihood $p(D \mid H_0)$ will equal the p-value computed for the voxel. In activated regions, that will be orders of magnitude lower than $p(D \mid H_a)$. One may therefore conclude that the likelihood ratio is high, and that H_a is strongly confirmed by the data.

This reasoning is mistaken, though. The problem lies in the move from a p-value to a low $p(D \mid H_0)$: there has been a tacit slide between two different, nonequivalent null hypotheses. The p-value at a voxel is the probability of seeing data like D if there was no BOLD response *at all*. The likelihood $p(D \mid H_0)$, on the other hand, is the probability of seeing D if the relevant voxel is *functionally unimportant*. The two will be equivalent only if alternative theories predict that functionally unimportant voxels show no differential BOLD response at all. That is, it must be the case that D would appear not just when H_a is correct, but *only* when H_a is correct. As Cacioppo et al. note, we rarely have evidence for the second half of that claim (Cacioppo et al. [2007]).

Consider, for example, the oft-cited work of Greene et al., which showed significant differences in activation in the angular gyrus when subjects were presented with emotionally laden moral dilemmas rather than impersonal ones (Greene et al. [2001]). Theories that attribute no role to emotion in moral decision-making need not be particularly threatened by these data. It is perfectly consistent with the alternative hypotheses that claim that the angular gyrus activation was part of a functionally unimportant reaction to the content of the moral dilemma. So, let D be the observed BOLD response in the angular gyrus, H_a be the hypothesis that the angular gyrus plays a functionally important role in moral reasoning, and H_0 the hypothesis that it plays no functionally important role. The likelihood $p(D \mid H_a)$ is high, of course: one would expect to see increased activity from a functionally important area. But $p(D \mid H_0)$ is also reasonably high: one would expect to see activation in the angular gyrus in response to emotionally laden scenarios, regardless of the functional role of the angular gyrus. Thus, the likelihood ratio is relatively low, and D does not provide especially compelling evidence. To generalize the argument, the poor temporal resolution of fMRI means that the evidence that a region is activated by a task can almost always instead be taken as an evidence that the region is activated as a functionally unimportant byproduct of task performance. As such, $p(D \mid H_0)$ will always be relatively high, and the likelihood ratio relatively low.

The problem of vague alternatives can manifest itself in various ways. Functional hypotheses rarely commit themselves to claims about activity in other, functionally unimportant areas. This means that $p(D \mid H_0)$ will be undefined in particular cases, and the strength of evidence impossible to determine (Mole and Klein [forthcoming]). This means that most experiments do not subject hypotheses to severe tests (Aktunç [unpublished]), and do not establish tight structure–function mappings (Cacioppo et al. [2007]). If alternative functional

hypotheses are simply indifferent about the response of functionally unimportant areas, then $p(D \mid H_0)$ may well be reasonably high, and at worst is undefined. So long as that is the case, even clean data need not be especially compelling.

4 Skepticism Is Due to NHST

The three forms of skepticism about neuroimages share a common root. SPMs present, at root, the results of numerous simultaneous NHSTs. That may seem like the least problematic fact about them. NHST is widely used in contemporary psychology. Those who attack the use of functional imaging typically do so in order to defend some *other* way of doing psychology, not because they are skeptical about psychology itself.[13] Yet NHST is theoretically controversial.[14] The controversy over NHST is unnecessarily polarized: I will assume that there are clear cases where NHST provides evidence, and clear cases where it does not. When we delineate the conditions in which NHST does not give good evidence, those conditions tend to obtain in cases where SPMs are used to test functional hypotheses.

First, NHST is uninformative for testing hypotheses about systems (or parts of systems) in which null hypotheses are usually false. This is a common complaint about the use of NHST in experimental psychology. Meehl, for example, notes that any null hypothesis of no effect must be false in dense systems, as there will always be minuscule but significant correlations between any two variables of interest (Meehl [1967], p. 108). Lykken similarly notes that, 'In psychology, everything is likely to be related at least a little bit to everything else, for complex and uninteresting reasons' (Lykken [1991], p. 31). Null hypotheses are rarely true in causally dense systems, making significance tests uninformative.[15] Causal density of the studied systems is a common feature of other disciplines in which NHST has been controversial.[16]

For the same reason, thresholding *p*-values in causally dense systems is also problematic. Causally dense systems typically show continuous variation in *p*-values, depending on the resolution of the test; picking a point at which a *p*-value changes from evidence against to evidence for a hypothesis is theoretically

[13] Landreth and Richardson, for example, argue that Uttal's skepticism reduces to skepticism about *t*-tests, which they consider a *reductio* (Landreth and Richardson [2004], p. 119).

[14] At the polemical extreme is Meehl's claim that NHST is 'basically unsound, poor scientific strategy, and one of the worst things that ever happened to the history of psychology.' (Meehl [1978], p. 817). For a recent review of the controversy, see (Nickerson [2000]); Morrison and Henkel ([1970]) and Harlow et al. ([1997]) collect many of the classic papers on the subject.

[15] In nondense systems the null hypothesis is not obviously false, because explanatory variables affect only a limited range of other variables. NHSTs may thereby provide useful information (Wainer [1999], p. 212; Kihlstrom [1998], p. 205; Hagen [1997], p. 20; Lewandowsky and Mayberry [1998], p. 210).

[16] See (McCloskey and Ziliak [1996]) in economics and (Johnson [1999]) in ecology.

arbitrary. As Abelson put it, 'We act foolishly when we celebrate results with $p = 0.05$ but wallow in self-pity when $p = 0.07$' (Abelson [1991], p. 12). In causally dense systems, again, no α can be compelling because false positives are extremely rare.

Finally, simple significance tests are most plausibly evidence for or against *ordinal* hypotheses: hypotheses that state that one parameter is larger than another (Frick [1996], p. 379). The evidence that univariate significance tests give is about the *direction*—positive, negative, or indeterminate—of an effect; they do not, on their own, provide information about the size of an effect or about the form of a relationship between variables of interest (Tukey [1991], p. 100). Functional hypotheses are not ordinal hypotheses, and building a functional theory requires more than information about the direction of differential performance (Newell [1973], p. 290). To say that some brain area A contributes to the performance of E is not merely to say *that* it does something when E is performed. Instead, it is a claim about *what* A does—namely, that it does something particular that contributes to the performance of E. It is perfectly possible for A to do something that makes a tiny but necessary contribution to E, or for A to be extremely active but have no effect on E.

These problems with NHSTs are, of course, just the problems outlined in Section 3. Skepticism about neuroimages is therefore simply a specific instance of a more general skepticism about NHST. NHST provides poor evidence for functional hypotheses about causally dense systems, in the brain or elsewhere.

5 Neuroimages versus Neuroimaging

Neuroimages are only one product of fMRI. Contemporary fMRI experiments produce more data than is summarized in neuroimages, and the data can be analyzed in a variety of ways. Can skepticism about neuroimages be generalized to skepticism about neuroimaging more generally?

I think not. First, most modern imaging experiments also present more sophisticated statistical analyses of fMRI data (Sarty [2007]). These more sophisticated analyses do not rely on the simplistic logic employed by NHSTs, and so are mostly immune from the critiques above. Second, information about neural anatomy can help constrain the implications of functional hypotheses in ways that permit more detailed testing. Experimenters likely do this in an informal way already when they choose regions of interest or interpret their data.[17] Information about neural connectivity can be formally integrated into experimental analysis via structural equation modeling and related techniques, and there is some reason to believe that the resulting evidence for functional

[17] See the work of Bartels et al., who argue that demonstrations of directional selectivity in V5/MT by fMRI always tacitly incorporate prior results from single-cell recordings (Bartels et al. [2008], p. 448).

hypotheses is more sensitive and compelling than that given by SPMs. Third, converging evidence from single-cell recordings and computational modeling can be used to give functional interpretations of quantitative measures of signal change (Logothetis [2008]; Bartels et al. [2008]).

The use of convergent information from modeling and anatomy is worthy of special note. Those who attack NHST typically argue that quantitative information is required to establish functional claims in causally dense systems. Quantitative information about the BOLD signal allows for more sophisticated hypotheses about the functional relationship between distinct brain regions, and is more likely to provide compelling evidence for functional hypotheses. In this regard, a comparison to structural MRI is telling. The production of structural MRIs is technically similar to the production of neuroimages;[18] the main difference is that structural MRIs show quantitative facts about the MR signal rather than the results of NHSTs. The images produced by structural MRI are routinely used to settle diagnostic disputes, and do not attract the general skepticism that attaches to fMRI.[19] Neuroimages are problematic because significance tests are not an adequate substitute for interpretable quantitative information.

Neuroimages are not worthless, however. The literature on NHST also allows us to offer an alternative interpretation of the role of significance tests, and so of neuroimages. On this alternative view, significance testing provides a first-pass sanity check on experimental data.[20] Finding statistical significance is never enough to *confirm* a hypothesis, but it does provide *warrant* for taking the data seriously and performing further analysis upon it. Similarly, neuroimages do not confirm functional hypothesis, but they do show brain areas in which the imaging data might be further used to confirm functional hypotheses. It requires more and different evidence to confirm functional hypotheses, but that evidence is something we can reasonably collect. This means that the production of neuroimages may be a necessary, though not sufficient, step in confirming functional hypotheses.

The evidence of neuroimaging is thus not in the images it produces. Nor do further data turn those images into evidence. Instead, neuroimages point us to where the evidence for functional hypotheses might be. Pictures of 'brain activity' are essentially uninterpretable without further analysis. Skepticism about neuroimages, however, does not provide grounds for general skepticism about the results of these further analyses.

[18] As far as the technology, principle, and most details of signal production are concerned, fMRI differs little from ordinary structural imaging. Early work on fMRI describes it simply as structural MRI that exploits deoxyhemoglobin as an endogenous contrast agent (Turner et al. [1993]).

[19] Which is not to say that there are no skeptics—see, e.g., (Joyce [2008]). Skepticism about structural MRI, however, tends to focus on its expense and its privileged role within the medical system, rather than on its status as evidence.

[20] See, for example, (Tukey [1969]; Abelson [1991]; Frick [1996]; Krueger [2001]).

Acknowledgements

Thanks to David Hilbert, Esther Klein, Chris Mole, Adina Roskies, Don Ross, audiences at University of Illinois at Chicago, and two anonymous reviewers for helpful discussion and comment.

1420 University Hall MC 267
University of Illinois at Chicago
601 S Morgan St
Chicago, IL 60607, USA
cvklein@uic.edu

References

Abelson, R. P. [1991]: 'On the Surprising Longevity of Flogged Horses: Why There Is a Case for the Significance Test', *Psychological Science*, **8**, pp. 12–5.

Aktunç, E. [unpublished]: 'Evidence and Hypotheses in Functional Neuroimaging'. Available online at: <emrah.aktunc.googlepages.com/AKTUNC.evid.hyp.fni.pdf>

Bartels, A., Logothetis, N. K. and Moutoussis, K. [2008]: 'fMRI and Its Interpretations: An Illustration on Directional Selectivity in Area V5/MT', *Trends in Neurosciences*, **31**, pp. 444–53.

Buxton, R. B. [2002]: *Introduction to Functional Magnetic Resonance Imaging: Principles and Techniques*, New York, NY: Cambridge University Press.

Cacioppo, J. T., Tassinary, L. G. and Berntson, G. G. [2007]: 'Psychophysiological Science: Interdisciplinary Approaches to Classic Questions about the Mind', in J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson (*eds*), *Handbook of Psychophysiology*, 3rd edition, New York, NY: Cambridge University Press, pp. 1–18.

Coltheart, M. [2006]: 'What Has Functional Neuroimaging Told Us about the Mind (So Far)?', *Cortex*, **42**, pp. 323–31.

Cummins, R. [1999]: 'Functional Analysis', in D. J. Buller (*ed.*), *Function, Selection, and Design*, Albany, NY: SUNY Press, pp. 57–84.

Dumit, J. [2004]: *Picturing Personhood: Brain Scans and Biomedical Identity*, Princeton, NJ: Princeton University Press.

Frick, R. W. [1996]: 'The Appropriate Use of Null Hypothesis Testing', *Psychological Methods*, **1**, pp. 379–90.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M. and Cohen, J. D. [2001]: 'An fMRI Investigation of Emotional Engagement in Moral Judgment', *Science*, **293**, pp. 2105–8.

Hagen, R. L. [1997]: 'In Praise of the Null Hypothesis Statistical Test', *American Psychologist*, **52**, pp. 15–24.

Hardcastle, V. G. and Stewart, C. M. [2002]: 'What Do Brain Data Really Show?', *Philosophy of Science*, **69**, pp. S72–82.

Harlow, L. L., Muliak, S. A. and Steiger, J. H. [1997]: *What If There Were No Significance Tests?*, Mahwah, NJ: Lawrence Erlbaum Associates.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L. and Pietrini, P. [2001]: 'Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex', *Science*, **293**, pp. 2425–30.

Huettel, S. A. and McCarthy, G. [2001]: 'The Effects of Single-Trial Averaging upon the Spatial Extent of fMRI Activation', *Neuroreport*, **12**, pp. 2411–6.

Huettel, S. A., Song, A. W. and McCarthy, G. [2004]: *Functional Magnetic Resonance Imaging*, Sunderland, MA: Sinauer Associates.

Jernigan, T. L., Gamst, A. C., Fennema-Notestine, C. and Ostergaard, A. L. [2003]: 'More "Mapping" in Brain Mapping: Statistical Comparison of Effects', *Human Brain Mapping*, **19**, pp. 90–5.

Johnson, D. H. [1999]: 'The Insignificance of Statistical Significance Testing', *Journal of Wildlife Management*, **63**, pp. 763–72.

Joyce, K. A. [2008]: *Magnetic Appeal: MRI and the Myth of Transparency*, Ithaca, NY: Cornell University Press.

Kihlstrom, J. F. [1998]: 'If You've Got an Effect, Test Its Significance; If You've Got a Weak Effect, Do a Meta-analysis', *Behavioral and Brain Sciences*, **21**, pp. 205–6.

Krueger, J. [2001]: 'Null Hypothesis Significance Testing: On the Survival of a Flawed Method', *American Psychologist*, **56**, pp. 16–26.

Landreth, A. and Richardson, R. C. [2004]: 'Localization and the New Phrenology: A Review Essay on William Uttal's *The New Phrenology'*, *Philosophical Psychology*, **17**, pp. 108–23.

Lewandowsky, S. and Mayberry, M. [1998]: 'The Critics Rebutted: A Pyrrhic Victory', *Behavioral and Brain Sciences*, **21**, pp. 210–1.

Lloyd, D. [2002]: 'Studying the Mind from the Inside Out', *Brain and Mind*, **3**, pp. 243–59.

Logothetis, N. K. [2008]: 'What We Can Do and What We Cannot Do with fMRI', *Nature*, **453**, pp. 869–78.

Lykken, D. T. [1991]: 'What's Wrong with Psychology, Anyway?', in D. Cicchetti and W. M. Grove (*eds*), *Thinking Clearly about Psychology: Essays in Honor of Paul E. Meehl*, Volume 1, Minneapolis, MN: University of Minnesota Press, pp. 3–39.

McCabe, D. P. and Castel, A. D. [2008]: 'Seeing Is Believing: The Effect of Brain Images on Judgments of Scientific Reasoning', *Cognition*, **107**, pp. 343–52.

McCloskey, D. N. and Ziliak, S. T. [1996]: 'The Standard Error of Regression', *Journal of Economic Literature*, **34**, pp. 97–114.

Meehl, P. E. [1967]: 'Theory-Testing in Psychology and Physics: A Methodological Paradox', *Philosophy of Science*, **34**, pp. 103–15.

Meehl, P. E. [1978]: 'Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology', *Journal of Consulting and Clinical Psychology*, **46**, pp. 806–34.

Mole, C. and Klein, C. [forthcoming]: 'Confirmation, Refutation and the Evidence of fMRI', in M. Bunzl and S. Hanson (*eds*), *Foundational Issues of Human Brain Mapping*, Cambridge: MIT Press.

Morrison, D. E. and Henkel, R. E. [1970]: *The Significance Test Controversy: A Reader*, Chicago, IL: Aldine Publishing.

Nair, D. G. [2005]: 'About Being BOLD', *Brain Research Reviews*, **50**, pp. 229–43.

Newell, A. [1973]: 'You Can't Play 20 Questions with Nature and Win', in W. Chase (*ed.*), *Visual Information Processing*, New York, NY: Academic Press, pp. 283–308.

Nickerson, R. S. [2000]: 'Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy', *Psychological Methods*, **5**, pp. 241–301.

Nir, Y., Dinstein, I., Malach, R. and Heeger, D. J. [2008]: 'BOLD and Spiking Activity', *Nature Neuroscience*, **11**, pp. 523–4.

Roskies, A. L. [2007]: 'Are Neuroimages like Photographs of the Brain?', *Philosophy of Science*, **74**, pp. 860–72.

Sarty, G. E. [2007]: *Computing Brain Activity Maps from fMRI Time-series Images*, Cambridge: Cambridge University Press.

Savoy, R. L. [2001]: 'History and Future Directions of Human Brain Mapping and Functional Imaging', *Acta Psychologica*, **107**, pp. 9–42.

Thirion, B., Pinel, P., Mériaux, S., Roche, A., Dehaene, S. and Poline, J.-B. [2007]: 'Analysis of a Large fMRI Cohort: Statistical and Methodological Issues for Group Analyses', *NeuroImage*, **35**, pp. 105–20.

Tukey, J. W. [1969]: 'Analyzing Data: Sanctification or Detective Work?', *American Psychologist*, **24**, pp. 83–91.

Tukey, J. W. [1991]: 'The Philosophy of Multiple Comparisons', *Statistical Science*, **6**, pp. 100–16.

Turner, R., Jezzard, P., Wen, H., Kwong, K., le Bihan, D., Zeffiro, T. and Balaban, R. [1993]: 'Functional Mapping of the Human Visual Cortex at 4 and 1.5 Tesla Using Deoxygenation Contrast EPI', *Magnetic Resonance in Medicine*, **29**, pp. 277–9.

Uttal, W. R. [2001]: *The New Phrenology*, Cambridge, MA: MIT Press.

Viswanathan, A. and Freeman, R. D. [2007]: 'Neurometabolic Coupling in Cerebral Cortex Reflects Synaptic More Than Spiking Activity', *Nature Neuroscience*, **10**, pp. 1308–12.

Viswanathan, A. and Freeman, R. D. [2008]: 'Reply to "BOLD and Spiking Activity"', *Nature Neuroscience*, **11**, p. 524.

Wainer, H. [1999]: 'One Cheer for Null Hypothesis Significance Testing', *Psychological Methods*, **6**, pp. 212–3.