Over-Interpreting Functional Neuroimages

Marc A. Burock

Cognitive neuroscientists use functional magnetic resonance imaging (fMRI) to measure properties of a participant's brain during a cognitive task. These imaging results are transformed into compelling pictures of brain activity using statistical models. I will argue that, for a broad class of experiments, neuroimaging experts have a tendency to over-interpret the functional significance of their data. This over-interpretation appears to follow from contentious theoretical assumptions about the mind-brain connection, and from a propensity to conflate the anatomical location of a statistically-significant correlation with knowledge of the mechanistic functioning at that location.

1. Introduction. Functional magnetic resonance imaging (fMRI) is a tool used by cognitive neuroscientists to create the compelling pictures of brain activity that most of us have come across in the public media and lay press. Although these pictures are taken as scientific facts, there is a growing concern within the fMRI community itself that the field lacks a degree of scientific rigor (Miller 2008). A recent controversy brought this concern to a head, which began when the *New York Times* published a column describing an fMRI study on undecided voters prior to the 2008 U.S. presidential election (Iacoboni et al. 2007). In this experiment, participants' brains were scanned as they viewed pictures and videos of the major presidential candidates at the time. The authors concluded, based largely upon patterns of measured brain activity, that "emotions about Hillary Clinton are mixed," "Mitt Romney shows potential," "John Edwards has promise—and a problem," and "Barack Obama and John McCain have work to do."

These claims created an immediate backlash within the fMRI community. Russell Poldrack, a cognitive neuroscientist who uses fMRI, said of the study "It epitomized everything that a lot of us feel is wrong about where certain parts of the field are going, which is throw someone in a scanner and tell a story about it" and that "people will start to see fMRI as neophrenology, just telling stories and not giving explanations."   (Miller 2008, 1412)

Within the philosophy of science, Roskies (2007) has considered the evidential status of fMRI images as compared to photographs. She concluded that proper interpretation of brain images involves far more than viewing a photograph, but suggests that neuroscientists familiar with fMRI would not misinterpret brain images in the ways she describes and that her findings are most relevant to the non-expert public. Racine et al. (2005) argues similarly that the public sees neuroimaging results without understanding the limitations and complexities of the work, which leads readily to misinterpretations. The public aside, fMRI experts, who presumably understand the methodology involved in creating functional brain images, can also misinterpret fMRI study results—at least according to Poldrack and others. While experts commonly disagree, it is not common for parts of a scientific community to feel or fear that its practitioners are heading toward pseudoscience and story-telling. There appears to be a fundamental conceptual problem at work.

In this paper I will attempt to analyze ways in which an fMRI expert may over-interpret her data. I will assume that some form of scientific induction is valid. By over-interpret I mean that she makes inferences that go beyond the quantitative meaning of a brain image, or draws inferences based upon the data in conjunction with contentious theoretical assumptions. I realize that this former sense of over-interpretation is vague and I will demonstrate it more clearly below. Put briefly, functional brain images directly represent mathematical constructs and not physical processes. When the quantitative nature of a brain image is carefully considered, I will argue that it says nothing directly about brain function.

With regard to the second sense of over-interpretation, I will focus upon an fMRI expert's philosophical assumptions about the mind-brain problem and show how these assumptions can influence scientific inferences. I expect an immediate reply that cognitive neuroscientists, as scientists, do not hold particular positions on the mind-brain problem or that they place these assumptions aside while following a scientific method. Perhaps that is true, yet cognitive neuroscientists do attempt to understand the mental by studying the brain in the first place—must they not consider cognition and brain related in some way a priori, even if that relation is a naïve reduction? I will not belabor this argument. While it is certainly possible that a cognitive neuroscientist holds no identifiable theoretical stance on the mind-brain problem, I will focus upon those who do.

Of course I will not be able to discover exactly what theories, if any, are being used by the fMRI community as a whole. Instead I will consider, as a model, one particular researcher who is prominent and respected in the fMRI community for her rigorous methods—Nancy Kanwisher. She is currently a professor of cognitive neuroscience at the Massachusetts Institute of Technology, a member of the National Academy of Sciences, and has used functional MRI for over a decade to study visual perception. She is anything but a straw man, and has even criticized other members of the fMRI community for their lack of rigor (Vul and Kanwisher, in press). I will attempt to understand her assumptions and inferences through her written work, focusing on her human fMRI experiments involving face-stimuli.

2. Technical Background. Before we can appreciate Kanwisher's philosophical assumptions and scientific assertions, it will be necessary to understand how functional brain images are created. There are two technical parts: (1) the biophysics of magnetic resonance imaging, and (2) the mathematical procedures that transform a 'raw' measurement into a picture. For the purposes of this paper, I will focus upon (2) because it concerns our argument, and only briefly address (1) as it does not and has been discussed before in the philosophy of science by Roskies (2007).

Some understanding of (1) is necessary before I proceed. The fMRI signal is a consequence of the magnetic properties of water protons as they interact with the magnetic properties of hemoglobin. Since oxygenated and deoxygenated hemoglobin in the blood have different magnetic properties, the fMRI signal will covary with the ratio of oxygenated to deoxygenated hemoglobin at each brain location, producing the so-called Blood Oxygen Level Dependent (BOLD) response. The oxygenation ratio is dependent upon the metabolic demands of cortical tissue, thus the BOLD signal is associated with neuronal activity, although this relationship is not completely understood (Logothetis and Wandell 2004).

Once the BOLD fMRI signal is measured, this data can be transformed into pictures. The mathematical techniques for creating functional brain images derive largely from statistical modeling methods (Friston et al. 1995). We begin with a raw dataset consisting of the BOLD fMRI signals measured at a discrete array of cuboid brain locations known as voxels. Each voxel is associated with a discrete time series of BOLD fMRI measurements taken at fixed intervals during the course of the experiment. The most common but certainly not the only way to analyze this dataset is by fitting each time series, voxel-by-voxel, with a linear model of the form $Y = X\beta + \varepsilon$, where Y is a column vector representing the measured BOLD times series at each voxel, X is an experimental design matrix with indicator variables representing different task or stimulus conditions, β is a vector of the unknown parameters to be estimated from the data, and ε represents an unknown noise term.

Again, Y is our BOLD measurement, X codifies our assumptions about the structure of the experiment, and β are the parameters that we are most interested in knowing, for these explain, in a condensed manner, the variability of Y. Of course we cannot know β exactly—we can only estimate it from the data Y. There are straightforward and mathematically rigorous ways of doing this, but we need not address these here. Let us assume that we can calculate an estimate of β.

In a two condition experiment, for instance, when one stimulus is a flashing checkerboard and the other is a dark screen, we would estimate two parameters $\beta = \{\beta_1 \ \ \beta_2\}$ at each voxel, corresponding to the two stimulus conditions. In the simplest case, β_1 and $\beta_2$would represent the mean BOLD signal for each condition of the experiment. We could generate an approximate map of ‘brain activity’ by visualizing the estimated difference between β_1 and $\beta_2$divided by the standard error of this difference (the t-statistic) at each voxel.

I have simplified things greatly, and have ignored the substantial technical hurdles involved in creating the reliable images that are presented to other scientists and the public. As well, there are many other mathematical methods and models that may be applied to help understand an fMRI dataset, but the modeling illustrated above is the most commonly used method by cognitive neuroscientist to make inferences about brain function.

For readers familiar with the concept of statistical correlation, there is a straightforward way to understand how fMRI pictures of brain activity can be created. Let us take $S(t)$ to be a so-called ‘dummy’ stimulus vector constructed as below, along with $Y(t)$ which is the BOLD fMRI time series measured during the experiment. Further suppose that we removed any ‘artifacts’ from $Y(t)$ such as those artifacts due to head motion.

$$S(t) = \left\{ \begin{array}{l} 1, \ Face\ present\ at\ time\ t \\ 0, \ Non\ face\ present\ at\ time\ t \end{array} \right\} \text{ and } Y(t) = BOLD\ signal$$

By calculating the linear correlation coefficient between $S(t)$ and $Y(t)$, we can create a rough statistical map of ‘activation.’ Larger correlations will be called brain activity, assuming the correlations meet our test for statistical significance. Again, I am not claiming that this is what researchers do—I only offer this example to explain, in greatly simplified terms, what fMRI activity represents from a quantitative perspective. It is useful because it helps us to understand the basis for many other complicated mathematical techniques used in brain imaging.

In most fMRI experiments, brain activity means that $Y(t)$ covaries with the stimulus condition. This covariance or correlation is identified using a particular model. I will often refer to fMRI brain activity as a correlation because correlation underlies the statistical tests of significance used to generate the brain images presented in scientific papers. The images do not represent physical measurements; they represent, roughly, reliable correlations between indexed stimulus conditions and physical measurements.

3. Theoretical assumptions of an expert. I will take ‘Functional imaging of human visual recognition’ by Kanwisher et al. [1996] as a starting point for this discussion. Kanwisher’s experiments are classic in that they are stimulus-response experiments where she determines, in advance, a set of well-defined stimuli to be presented to the subject while measuring a neurophysiological property of that subject. I think that these types of experiments are the least susceptible to interpretational difficulties and that Kanwisher’s are particularly lucid in that her stimuli are visual images.

The fMRI experiments of interest purportedly study human visual recognition proper, which according to Kanwisher et al. ”include[s] two main components: (i) the high-level perceptual analysis of a visual stimulus (for example, the construction of a structural description of an object's shape), and (ii) the matching of that perceptual description to a stored visual representation in memory (e.g., determining that the shape looks more like a chair than a bicycle). Note that this definition excludes both very 'early' visual processes such as edge extraction, and very 'late' processes such as accessing the name or meaning of a visually-presented stimulus." (1996, 55)

In part (i) we are told that visual recognition includes perceptual analysis exemplified by the construction of a structural description of an object's shape, and in (ii) the matching of that perceptual description to a stored visual representation. I am interested in the actions mentioned in this definition, specifically 'perceptual analysis', 'construction', and 'matching'; and more urgently, who or what performs these actions. The subject of the action is left blank, and we wonder whether it is the human participant as an agent who performs these actions, or if Kanwisher implies that the brain or specific brain regions perform these actions.

From other published literature, it becomes clear that Kanwisher holds that particular brain regions or collections of regions perform these actions. She says, for instance, "…by demonstrating the extreme specificity of one cortical region for a single high-level function—face perception.", "…fMRI has revealed a particular region in the human brain where this special face perception machinery apparently resides…", "More generally, which functions get their own dedicated patch of cortex, and why?", and "Thus, Tsao et al. provide the strongest evidence yet for extreme specificity of a cortical region for a complex high-level function." (Kanwisher 2006, 617)

We see two allusions to the high-level function mentioned in the previous article. It appears Kanwisher believes that brain tissue analyzes perceptions, constructs structural descriptions, and matches perceptual descriptions to stored visual representations. If this were *only* a metaphoric way to express a research finding, then I would have no quarrel; however, these sorts of comments are prolific in the neuroimaging literature, and more, they are used in the rigorous definitions that circumscribe particular sub-domains of cognitive neuroscience.

On the surface, attributing human activities, such as matching, to specific cortical regions sounds like anthropomorphism. Human beings analyze perceptions, construct descriptions, and match descriptions to representations; neurons ‘maintain’ a potential difference and ‘generate’ action potentials. Within a purely physical way of talking about things, it should not be said that brain tissue analyzes anything, for brain tissue does nothing but evolve its physical properties according to the laws of physics. There may of course be localized ‘brain laws’, such as Kirkoff’s circuit laws for electronic circuits, but these local laws deal with the physical properties of circuits and have little to do with agent-based functioning.

I believe that Kanwisher would deny the allegation of anthropomorphism of brain tissue, yet in the quotations above, she is in some way ‘placing’ human functions ‘into’ pieces of brain tissue. For instance, we might conclude that face perception occurs ‘within’ a particular piece of brain tissue, or that a particular patch of tissue is solely responsible for face perception in-itself without any connection to other parts of the brain. I do not believe that many cognitive neuroscientists would agree with these conclusions, but they follow from a casual reading of the above quotations.

Again, I am not denying the methodological rigor of Kanwisher’s fMRI research—which I believe is top-notch—but I am questioning her underlying assumptions about the mind-brain problem, and using her way of talking about her research as a model for how the field discusses functional brain-imaging results as a whole. Even if she does not intend my interpretation, her language can at least be construed as form of anthropomorphism of brain tissue. She appears to be placing human characteristics into clumps of cells.

While I see anthropomorphism at work here, I realize that this would not be the common philosophical interpretation. Given the quotes above, most philosophers would likely attribute cognitivism to Kanwisher’s theoretical stance in the general sense that she takes the brain to be similar to a digital computer, or that cognition is the product of algorithm-like rules run on the brain machine.[1] Her specific comment about special face-perception machinery, localized to particular cortical areas, makes this position seem clear, and her broader use of a computational language supports it.

[1] Cognitivism is two-stage anthropomorphism. Human beings *constructively* placed human functions—like mathematical calculation—within computers. To take the computer algorithms and place them, metaphorically, into brain tissue is a form of anthropomorphism if we consider the human design of computers at the start.

Kanwisher, if pressed for a philosophical stance on the mind-brain problem, may say that as a scientist she is objectively neutral on the issue, but her language betrays a particular theoretical orientation that many philosophers will readily identify. She may maintain that her language is metaphorical and has nothing to do with her science. However, I will argue that her philosophical orientation to the mind-brain problem, which appears to be some version of cognitivism, influences her scientific inferences and leads her to over-interpret the data. I will speculate that she is hardly alone.

4. Evaluating functional inferences. A scientifically interesting stimulus-response fMRI experiment demonstrates cortical brain regions that selectively respond to specific stimuli. By this I mean, a cortical region is discovered to be relatively active if and only if a specific stimulus or class of stimuli are presented. The if and only if clause is never established empirically, thus the researcher is content to test a pre-chosen set of stimuli that contain a fair degree of variation. Kanwisher does this for face stimuli in her 1997 paper, discovering a region of cortex that is a "module in extrastriate cortex specialized for face perception" (1997, 4302) where face perception involves detecting and representing faces. This cortical area has been labeled the fusiform face area (FFA).

While I agree that the activity in the FFA is selectively correlated, in a statistical sense, with presented face images, it is not clear that we can infer any other facts from this observation without appealing to fundamental assumptions about the brain-mind connection. To make this concern solid, I will consider several functional inferences that may be endorsed given a rigorously performed stimulus-response fMRI experiment. I have chosen the inferences below for a variety of reasons. The first inference I find logically tractable, the last is something I cannot refute, and the remaining are inferences likely to be endorsed by Kanwisher and others within the fMRI community. Many other inferences are possible, but the following, taken as whole, demarcate my areas of concern.

1. The activity of FFA neurons is necessary and/or sufficient for the perception of faces.

I have not seen (1) explicitly endorsed by Kanwisher, but it is a tractable statement she might defend, especially since she identifies the FFA as the locus of face perception. Yet if the only evidence we have available derives from a stimulus-response fMRI experiment, then we have no ground to defend (1). This data does not logically determine the necessary and sufficient conditions for the perception of faces, for it is possible that a subject without the correlations identified by Kanwisher perceives faces just fine, and that someone with these FFA correlations cannot perceive faces at all. Her experiment offers nothing to rule-out these possibilities, and thus (1) cannot be endorsed on the basis of a stimulus-response fMRI experiment alone, at least not until some greater scientific fact connects these correlations to necessary and sufficient conditions.

2. The activity of FFA neurons represents the face stimulus.

There is some evidence that Kanwisher takes this position to be true. In demarcating the scientific study of visual recognition, she gives as an example the high-level function, "construction of a structural description of an object's shape," (1996, 55) and states elsewhere that high-level functions occur within a patch of cortical tissue. A structural description of an object is a representation of that object. Therefore, one might conclude that the neuronal activity in the FFA represents faces. But to whom does the FFA re-present? Does it represent it to the experimenter? Does it represent a face to the subject's brain as a whole? If so, what does that mean?

There are those who study neuronal activity at the level of action potentials who have rigorously considered the representational *capacity* of neurons. An excellent survey of this research and its results can be found in the book *Spikes* by Rieke et al. [1999]. These authors consider the temporal sequence of neuronal action potentials (spike trains), and how they *can be* quantitatively related to stimuli. Unlike many cognitive neuroscientists, these authors explicitly state their theoretical assumptions about the mind-brain problem, explaining that they are taking a homuncular view of representation as a starting point for mathematical analysis. This does not detract from their results. They are in no way claiming, as a scientific fact, that neuronal activity represents perceptual stimuli. Rather, they ask, what is the representational capacity of neuronal activity given a homuncular perspective?

In contrast, the selective activity in an fMRI experiment does not, by itself, establish that a particular cortical region represents anything, for the mathematics involved in computing selective correlations and the statistical models that identify 'active' brain regions have nothing to do with analyzing the structure of the fMRI signal for its representational capacity. Kanwisher may attempt these calculations for fMRI time series, but she has not done so to my knowledge.

Further, representational capacity and the act of representing are two different things. The former is quantitatively tractable within the neurosciences while the latter currently is not. Even if localized activity did represent the stimulus in question, we would still have to consider the meaning of this representation from a mind-brain perspective. I imagine that many cognitive neuroscientists would say that the activity represents the stimulus to the brain, implicitly equating the whole brain with the homunculus. One may of course elaborate a specific theory that attempts to eliminate the homunculus altogether, and with it, the idea that neuronal activity represents stimuli, but then we would agree that (2) is an unjustified scientific inference.

An analogy within the field of dendrochronology—the science of tree-ring dating—may be helpful. While tree-rings represent the age of the tree to an observing dendrochronologist, it is another thing to say that the tree-rings represent the age of the tree to the tree itself. Rather, the tree appears quite oblivious to its number of rings and the concept of number altogether. Although a brain may possess an unspecified self-representational capacity, it is far from certain that the physical matter of the brain can or does interpret its own physical properties as representations.

A cognitive neuroscientist has no ground to endorse (2) given only the selective activity derived from an fMRI stimulus-response experiment, for these data do not address the problem of representation at all. More, any claim of representing almost certainly involves a particular philosophical stance on the mind-brain problem, thus the scientist cannot claim to be objectivity neutral to the mind-brain connection and still endorse (2).

3. The activity of FFA neurons indicates when a face is being perceived.

The homuncular problem arises once again. Indicates to whom? The brain? The experimenter? I am not denying that the selective activity found in an fMRI experiment is useful. There are ways of using this data for practical purposes. For instance, given a pattern of fMRI activity for a single subject, we may be able to predict the stimulus that accompanied that activity. This is a fun trick. Langleben et al. [2005] used it for lie detection. As well, given a stimulus, one may predict the pattern of fMRI activity as Mitchell et al. [2008] did for pictured nouns. These pragmatic uses of fMRI activity are technologically exciting, but they are also completely blind to the mind-brain problem, the nature of the brain, and the nature of the mind. In other words, our ability to make these predictions does not require that we understand the structure or function of the mind or brain.

Predicative uses of fMRI activity are performed using the mathematical tools discovered in the field of pattern classification (see Webb 2002 for an intro to these tools). So long as the variability in stimuli is correlated with the variability of fMRI signals, we will be able to use classification algorithms to relate the two. To use these algorithms, nothing need be known—or assumed—about the mind or brain. The measured activity becomes a meaningless vector of numbers associated with an indexed class of stimuli. A computer scientist need not even know what the numbers represent in order to perform these classifications. If the vectors statistically differ by class, she will be able to perform these neat predications. While the activity of cortical neurons may statistically indicate to an outside observer when a particular stimulus is present, that indication does not imply that the activity plays the same role within the perceiver. This role requires an inner homunculus to view these patterns and to signal an alarm. The statement in (3) may be endorsed with the following changes: These cortical neurons (probabilistically) indicate (to an outside observer) when a face stimulus was presented to the observing subject.

4. FFA neurons perform a high-level function (face perception).

Kanwisher does make statements nearly identical to (4). But neurons do not perceive faces, people perceive faces. And her data is silent as to the mechanistic functioning of the FFA—she has only suggested *where* a hypothetical mechanism might exist, if it exists at all. Nor would the FFA alone make the perception of faces possible, and her experimental data cannot establish that the FFA is even necessary for face perception.

When supporting (4), I believe that Kanwisher would make the following argument based upon the results of a functional imaging experiment:

Data : Simultaneous BOLD fMRI signals and face/non-face stimuli

Statistical Inference **:** The FFA is a region where the BOLD signals and stimuli are selectively correlated

Functional Inference**:** The FFA performs face perception (because the BOLD signals and stimuli are selectively correlated within the FFA)

But the functional inference is circular because the FFA is defined as the region where face stimuli and signals are correlated. To refute the circularity of the functional inference, Kanwisher must use evidence other than the correlations or appeal to assumptions outside of the data. Since fMRI experiments of this type provide no other evidence, the functional inference must be based upon an unspecified assumption. Kanwisher is simply making a circular or speculative inference when she endorses something like (4) given her findings of selective fMRI correlations.

5. FFA neurons 'process' or 'convey' face information.

It is quite in vogue, within the fMRI community and the field of cognitive neuroscience, to say that particular cortical areas process different kinds of information. For instance, Kanwisher might argue that the FFA processes facial information. But what exactly do she and other neuroscientists mean when they make such claims? It is well known that "information is notoriously a polymorphic phenomenon and polysemantic concept" (Floridi 2008), so a scientific fact couched in the language of information, without further specification, is ambiguous at best.

Again, we can contrast this casual use of information in (5) with that used by the authors of *Spikes*. They use Claude Shannon's (1948) mathematical theory of communication (aka mathematical information theory) to estimate the information transmission-rate of a spike train about a stimulus. Like pattern classification algorithms, these methods are blind to the semantic or biologically relevant meaning of a spike-train. A thorough mathematical explanation of communication theory is beyond the scope of this work. Briefly, imagine that an outside stimulus is communicated to the biological organism in the form of a spike train. Given noisy communication between environment and organism, receiving a message (a spike train) reduces the organism's uncertainty about the transmitted message (the stimulus). At best, receiving a particular spike train will tell it exactly what stimulus was sent. At worst, reception of the spike train tells it nothing about the stimulus.

Rieke et al. derive, with clearly specified assumptions, theoretical bounds on the informational carrying capacity of spike trains, and show that these bounds are nearly met in some situations. From the perspective of a homunculus or outside observer, we may say that spike trains have the capacity to communicate information about the environment, and near optimally so—a result I find fascinating. However, this result does not imply that we understand a brain mechanism; rather, Rieke et al. have identified a capacity given a homuncular perspective. When we take a homuncular perspective and use the word information in Shannon's quantitative sense, we can unambiguously claim that cortical neurons have the capacity to convey information. Few authors of neuroimaging experiments explicitly state this position or make use of mathematical communication theory.

Although I accept Rieke's notion of information transmission applied to the neurosciences, does selective activity in a neuroimaging experiment demonstrate that specific patches of brain tissue 'process' specific kinds of information? Can we say that the FFA processes face information? In a very general way we may say this, in the sense that the FFA receives spike train inputs and produces spike train outputs. Yet every region of the brain presumably processes face information in this sense. Primary visual cortex has inputs and produces outputs to face stimuli. Kanwisher may point out that the FFA selectively responds to face stimuli, whereas primary visual cortex is non-selective. She would like to infer that the FFA processes face and only face information, but the selective activity of an fMRI experiment cannot, by itself, establish this fact. I say this because we have no idea what the FFA is doing—we only know that BOLD measurable activity is correlated with the presentation of face/non-face signals. If one takes selective correlations to be equivalent with "face information processing," then I have no argument. If, however, “face information processing” is a scientific fact beyond a measurable correlation, then we would like to know what additional assumptions are needed to establish this fact.

Given the ambiguous meaning of information, it is difficult to know what a scientist is talking about when she uses this word to describe her scientific facts. In Kanwisher et al. (1996) there is evidence that she equates visual information with the various quantifiable ways that one may describe a visual object. For example, visual information can be the spatial relations between features, the orientation of the object, the color of the object, the luminance, the outline or shape, the complexity, or the spatial frequency content. I believe that attributing face information processing to the FFA, in Kanwisher’s sense, is similar to saying that the FFA calculates, or represents, the quantifiable properties of a face image. But an fMRI stimulus-response study does not demonstrate any mechanism of calculation, and the problem of representation was discussed in (3).

6. The activity of FFA neurons is selectively correlated with face stimuli, but is functionally irrelevant.

In an fMRI stimulus-response experiment, where the inferences follow from correlations between stimuli and BOLD signals, nothing rules out the possibility that those correlations are accidental—not in the sense that the correlations are statistically spurious, but that those correlations are functionally irrelevant to the stimuli of interest. As an analogy, suppose my computer has a CPU fan with a blue LED light on the fan. The light, however, is unlit and the fan isn’t spinning. It happens that when I kick my computer just so on the left side of the front cover, the LED lights up, the fan begins spinning but stops after a second or two, and the light goes out. If I kick it again, just so, it starts up for a second then stops. I can reliably cause the fan to turn on for a bit. When I kick the computer in other places, or shake it up, or sing to it, nothing happens to the fan. The fan is selectively correlated with a specific kick. Perhaps there are hundreds of computers, constructed at the same factory, that behave similarly. This apparently causal relationship does not imply that the fan is functionally relevant to my kicking, or processes kicking information, or represents kicking. In fact, my kicking and the fan are, as

Aristotle might say, accidentally related. Given *only* the selective correlations of a neuroimaging study, it is possible that the FFA is functionally irrelevant with respect to face stimuli.

5. Concluding remarks. A subsequent lesion study has shown that FFA damage is associated with an impaired ability to make facial judgments (Barton et al. 2002), while other fMRI studies challenge the idea that the FFA selectively responds to faces at all (Gauthier et al. 2000) and demonstrate that FFA activation is not sufficient for face perception (Hasson et al. 2003). Whatever future research may come, Kanwisher's scientific fact—that the BOLD fMRI signal in the FFA is relatively greater when face stimuli are presented as compared to a variety of non-face stimuli—will remain intact. This fact has already guided subsequent research. It does not follow that we understand anything about a brain mechanism from this fact or that we are justified in making inferences 1-5 as above, at least not without dogmatically asserting a particular stance on the mind-brain connection or using ambiguous language.

I believe that Kanwisher and many others within the fMRI community assume some form of cognitivism (which is unlikely to be a great surprise). This assumption, however, is philosophically contentious and scientifically unneeded, and more worrisome, invites the use of a metaphorical computational language to describe neuroimaging findings, leading to over-interpretation. I have also seen evidence that Kanwisher implicitly takes the whole brain to be a homunculus of sorts without specifying what this means. Racine et al. (2005) observed a similar trend in the popular media, calling this practice 'neuro-essentialism,' which they define as equating subjectivity and personal identity to the brain. I have called this a form of anthropomorphism.

Perhaps the primary source of over-interpretation follows from Kanwisher's tendency to conflate the anatomical location of a statistically-significant correlation with knowledge of the mechanistic functioning at that location. But Kanwisher's use of fMRI data is mechanistically and functionally silent. I say this because after she infers the locations of significant correlations and states precisely what two things are being correlated, she sets the data aside and focuses upon the derived pictures. She does not use the data to further our understanding of a mechanism or function; rather, she infers that the correlation-selected location performs a mechanistic function because stimuli and fMRI BOLD signals were correlated at that location. Put this way, the reasoning appears circular.

I chose to analyze Kanwisher's fMRI face-stimuli experiments because her subject matter, the visual face, is relatively well-defined and less open to interpretational difficulties. Contrast her work with, for example, a paper that used fMRI to investigate emotions while making moral judgments (Greene et al., 2001). In this work the fMRI BOLD signal was measured on participants as they read 60 'practical dilemmas' that were divided into moral and non-moral categories, during which the participants judged the 'appropriateness' of a suggested action one might perform in the given scenario. Finding selectively-correlated differences for categories and judgments, the authors concluded that the spectrum of moral dilemmas differentially engages emotional processing and that this finding illuminates issues in moral philosophy. Although this work was published in the prestigious journal *Science*, once we see the interpretational difficulties involved with Kanwisher's lucid experiments, we cannot help but feel that Greene et al. is telling a story that goes far beyond the measured correlations.

References

Barton, J.J, D.Z. Press, J.P Keenan, and M. O'Connor (2002), "Lesions of the Fusiform Face Area Impair Perception of Facial Configuration in Prosopagnosia", *Neurology* 58: 71-78.

Floridi, Luciano, (Winter 2008 Edition) "Semantic Conceptions of Information", *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2008/entries/information-semantic/>.

Friston, K. J., A. P. Holmes, J-B. Poline, P. J. Grasby, S. C. R. Williams, R. S. J. Frackowiak, and R. Turner (1995), "Analysis of fMRI Time-Series Revisted", *Neuroimage* 2: 45-53.

Gauthier, Isabel, Pawel Skudlarski, John C. Gore, and Adam W. Anderson (2000), “Expertise for Cars and Birds Recruits Brain Areas Involved in Face Recognition”, *Nature Neuroscience* 3: 191-197.

Greene, Joshua D., R. Brian Sommerville, Leigh E. Nystrom, John M. Darley, and Jonathan D. Cohen (2001), “An fMRI Investigation of Emotional Engagement in Moral Judgment”, *Science* 293: 2105-2108.

Hasson, Uri, Galia Avidan, Leon Y. Deouell, Shlomo Bentin, and Rafael Malach (2003), “Face-Selective Activation in a Congenital Prosopagnosic Subject”, *Journal of Cognitive Neuroscience* 15(3): 419-431.

Iacoboni, Marco, Joshua Freedman, Jonas Kaplan, Kathleen H. Jamieson, Tom Freedman, Bill Knapp and Kathryn Fitzgerald (2007), “This Is Your Brain on Politics”, *The New York Times*, November 11, A14.

Kanwisher, Nancy, Marvin M. Chun, Josh McDermott, and Patrick J. Ledden (1996), “Functional Imaging of Human Visual Recognition”, *Cognitive Brain Research* 5: 55-67.

Kanwisher, Nancy, Josh McDermott, and Marvin M. Chun (1997), “The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception”, *The Journal of Neuroscience* 17: 4302-4311.

Kanwisher, Nancy and Galit Yovel (2006), “The Fusiform Face Area: A Cortical Region Specialized for the Perception of Faces”, *Philosophical Transactions of the Royal Society B* 361: 2109-2128.

Kanwisher, Nancy (2006), “What’s in a Face?”, *Science* 311: 617-618.

Langleben , Daniel D., James W. Loughead, Warren B. Bilker, Kosha Ruparel, Anna Rose Childress, Samantha I. Busch, and Ruben C. Gur (2005), “Telling Truth From Lie in Individual Subjects with Fast Event-Related fMRI”, *Human Brain Mapping* 26(4):262-72.

Logothetis, Nikos K., and Brain A. Wandell (2004), “Interpreting the BOLD Signal”, *Annual Review of Physiology* 66: 735-769.

Logothetis, Nikos K. (2008), “What We Can Do and What We Cannot Do with fMRI”, *Nature* 453: 869-878.

Miller, Greg (2008), “Growing Pains for fMRI”, *Science* 320: 1412-1414

Mitchell, Tom M., Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang,Vicente L. Malave,Robert A. Mason, and Marcel Adam Just (2008), "Predicting Human Brain Activity Associated with the Meanings of Nouns", *Science* 320: 1191-1195.

Racine, Eric, Ofek Bar-Ilan, and Judy Illes (2005), "fMRI in the Public Eye", *Nature Reviews Neuroscience* 6: 9-14.

Rieke, Fred, David Warland, Rob de Ruyter van Steveninck, and William Bialek (1999), *Spikes: Exploring the Neural Code*. Cambridge, MA: MIT Press.

Roskies, Adina L. (2007), "Are Neuroimages Like Photographs of the Brain?", *Philosophy of Science* 74: 860-872.

Searle, John (1992), *The Rediscovery of the Mind.* Cambridge, MA: MIT Press

Shannon, Claude E. (1948), "A Mathematical Theory of Communication", *The Bell System Technical Journal* 27: 379-423, 623-656.

Vul, E. and Nancy Kanwisher (in press) "Begging the Question: The Non-independence Error in fMRI Data Analysis" To appear in Hanson, S. and M Bunzl (Eds.) *Foundations and Philosophy for Neuroimaging*.

Webb, Andrew (2002), *Statistical Pattern Recognition*, 2nd edition. Malvern, UK: John Wiley and Sons Ltd.