Inferring Mental States from Neuroimaging Data: From Reverse Inference to Large-Scale Decoding

Russell A. Poldrack[1,*]
[1]Imaging Research Center and Departments of Psychology and Neurobiology, University of Texas at Austin, 3925-B W. Braker Lane, Austin, TX 78759, USA
*Correspondence: poldrack@mail.utexas.edu

DOI 10.1016/j.neuron.2011.11.001

A common goal of neuroimaging research is to use imaging data to identify the mental processes that are engaged when a subject performs a mental task. The use of reasoning from activation to mental functions, known as "reverse inference," has been previously criticized on the basis that it does not take into account how selectively the area is activated by the mental process in question. In this Perspective, I outline the critique of informal reverse inference and describe a number of new developments that provide the ability to more formally test the predictive power of neuroimaging data.

Understanding the relationship between psychological processes and brain function, the ultimate goal of cognitive neuroscience, is made particularly difficult by the fact that psychological processes are poorly defined and not directly observable, and human brain function can only be measured through the highly blurred and distorted lens of neuroimaging techniques. However, the development of functional magnetic resonance imaging (fMRI) 20 years ago afforded a new and much more powerful way to address this question in comparison to previous methods, and the fruits of this technology are apparent in the astounding number of publications using fMRI in recent years.

The classic strategy employed by neuroimaging researchers (established most notably by Petersen, Posner, Fox, and Raichle in their early work using positron emission tomography; Petersen et al., 1988; Posner et al., 1988) has been to manipulate a specific psychological function and identify the localized effects of that manipulation on brain activity. This has been referred to as "forward inference" (Henson, 2005) and is the basis for a large body of knowledge that has been derived from neuroimaging research. However, since the early days of neuroimaging, there has also been a desire to reason backward from patterns of activation to infer the engagement of specific mental processes. This has been called "reverse inference" (Poldrack, 2006; Aguirre, 2003) and often forms much of the reasoning observed in the discussion section of neuroimaging papers (under the guise of "interpreting the results"). In some cases, reverse inference underlies the central conclusion of a paper. For example, Takahashi et al. (2009) examined the neural correlates of the experience of envy and schadenfreude. They found that envy was associated with activation in the anterior cingulate cortex, in which they note, "Cognitive conflicts or social pain are processed" (p. 938), whereas schadenfreude was associated with activation in the ventral striatum, "a central node of reward processing" (p. 938). The abstract concludes as follows: "Our findings document mechanisms of painful emotion, envy, and a rewarding reaction, schadenfreude," in which the psychological states (i.e., pain or reward) are inferred primarily from activation in specific regions (anterior cingulate or ventral striatum). This is just one of many examples of reverse inference that are evident in the neuroimaging literature, and even the present author is not immune.

Reverse inference is also common in public presentations of imaging research. A prime example occurred during the US Presidential Primary elections in 2007, when the *New York Times* published an op-ed by a group of researchers titled "This is Your Brain on Politics" (Iacoboni et al., 2007). This piece reported an unpublished study of potential swing voters who were shown a set of videos of the candidates while being scanned using fMRI. Based on these imaging data, the authors made a number of claims about the voters' feelings regarding the candidates. For example, "When our subjects viewed photos of Mr. Thompson, we saw activity in the superior temporal sulcus and the inferior frontal cortex, both areas involved in empathy," and, "Looking at photos of Mitt Romney led to activity in the amygdala, a brain area linked to anxiety." More recently, another *New York Times* op-ed by a marketing writer used unpublished fMRI data to infer that people are "in love" with their iPhones (Lindstrom, 2011). Clearly, the desire to "read minds" using neuroimaging is strong.

In 2006, I published a paper that challenged the common use of reverse inference in the neuroimaging literature (Poldrack, 2006; for a similar earlier critique, see Aguirre, 2003). Since the publication of those critiques, "reverse inference" has gradually become a bad word in some quarters, though very often a citation to those papers is used as a fig leaf to excuse the use of reverse inference. At the same time, a number of researchers have argued that it is a fundamentally important research tool, especially in areas such as neuroeconomics and social neuroscience, in which the underlying mental processes may be less well understood (e.g., Young and Saxe, 2009). In what follows, I will lay out and update the argument against reverse inference as it is often practiced in the literature. I will then describe how recent developments in statistical analysis and informatics have provided new and more powerful ways to infer mental states from neuroimaging data and discuss the limitations of those techniques. I will conclude by highlighting what I see as important challenges that remain in the quest to reliably use neuroimaging data to understand mental function.

A Probabilistic Framework for Inference in Neuroimaging

The goal of reverse inference is to infer the likelihood of a particular mental process M from a pattern of brain activity A, which can be framed as a conditional probability P(M|A) (see Sarter et al., 1996 for a similar formulation). Neuroimaging data provide information regarding the likelihood of that pattern of activation given the engagement of the mental process, P(A|M); this could be activation in a specific region or a specific pattern of activity across multiple regions. The amount of evidence that is obtained for a prediction of mental process engagement from activation can be estimated using Bayes' rule:

$$P(M|A) = \frac{P(A|M) \times P(M)}{P(A|M) \times P(M) + P(A|\sim M) \times P(\sim M)}$$

Notably, estimation of this quantity requires knowledge of the base rate of activation A, as well as a prior estimate of the probability of engagement of mental process M. Given these, we can obtain an estimate of how likely the mental process is given the pattern of activation. The amount of additional evidence that the pattern of activity provides for engagement of the mental process can be framed in terms of the ratio between the posterior odds and prior odds, known as the Bayes factor. To the degree that the base rate of activation in the region is high (i.e., it is activated for many different mental processes), then activation in that region will provide little added evidence for engagement of a specific mental process; conversely, if that region is very specifically activated by a particular mental process, then activation provides a great deal of evidence for engagement of the mental process.

This framework highlights the importance of base rates of activation for quantifying the strength of any reverse inference, but such base rates were not easy to obtain until recently. In Poldrack (2006), I used the BrainMap database to obtain estimates of activation likelihoods and base rates for one particular reverse inference (viz., that activation of Broca's area implied engagement of language function). This analysis showed that activation in this region provided limited additional evidence for engagement of language function. For example, if one started with a prior of P(M) = 0.5, activation in Broca's area increased the likelihood to 0.69, which equates to a Bayes factor of 2.3; Bayes factors below 4 are considered weak. Others have since published similar analyses that were somewhat more promising; for example, Ariely and Berns (2010) found that activation in the ventral striatum increased the likelihood of reward by a Bayes factor of 9, which is considered moderately strong.

One drawback of the BrainMap database is that the papers in the database are manually chosen to be entered and thus reflect a biased sample of the literature. In recent work, we (Yarkoni et al., 2011) developed an automated means to obtain activation coordinate data (like those contained in BrainMap) from the full text of published articles; currently, the database contains data from 3,489 articles from 17 different journals. These data (which are available online at http://www.neurosynth.org) provide a less-biased means to quantify base rates of activation (though biases clearly remain due to the lack of complete and equal coverage of all possible mental states in the literature). Figure 1 shows a rendering of base rates of activation across the studies in this database. What is striking is the degree to which some of the regions that are the most common targets of informal reverse inference (e.g., anterior cingulate and anterior insula) have the highest base rates and therefore are the least able to support strong reverse inferences.

Reverse Inference using Literature Mining

A thorough analysis of reverse inference using meta-analytic data is difficult because it requires manual annotation of each data set in order to specify which mental processes are engaged by the task. Databases such as BrainMap rely upon relatively coarse ontologies of mental function, which means that although one can assess the strength of inferences for broad concepts such as "language," it is not possible to perform these analyses for finer-grained concepts that are likely to be of greater interest to many researchers.

An alternative approach relies upon the assumption that the words used in a paper should bear a systematic relation to the concepts that are being examined. Yarkoni et al. (2011) used the automatically extracted activation coordinates for 3,489 published articles, along with the full text of those articles, to test this form of reverse inference: instead of asking how predictive an activation map is for some particular mental process (as manually annotated by an expert), this analysis asked how well one can predict the presence of a particular term in the paper given activation in a particular region. Although there are clearly a number of reasons why this approach might fail, Yarkoni et al. (2011) found that for many terms it was possible to accurately predict activation in specific regions given the presence of the term (i.e., forward inference), as well as to predict the likelihood of the term in the paper given activation in a specific region (i.e., reverse inference). We also found that it was possible to classify data from individual participants with reasonable accuracy, as well as to classify the presence of words in individual studies against as many as ten alternatives, which suggests that these meta-analytic data can provide the basis for relatively large-scale generalizable reverse inference.

A challenge to the use of literature mining to perform reverse inference is that it is based on the language that researchers use in their papers and may thus tend to reify informal reverse inferences. For example, if researchers in the past tended to interpret activation in the anterior cingulate cortex as reflecting "conflict" based on informal reverse inference, then this will increase the support obtained from a literature-based meta-analysis for this reverse inference (because that analysis examines the degree to which the presence of activation in the anterior cingulate is uniquely predictive of the term "conflict" appearing in the text). Another challenge for this approach arises from the coarse nature of coordinate-based meta-analytic data, which will probably limit accurate generalization to domains in which the relevant activation is distributed across large areas rather than being reflected in finer-grained patterns of activation; for example, it will be much easier to identify data sets in which visual motion is present than to identify a particular motion direction. Finally, literature-based analysis is complicated by the

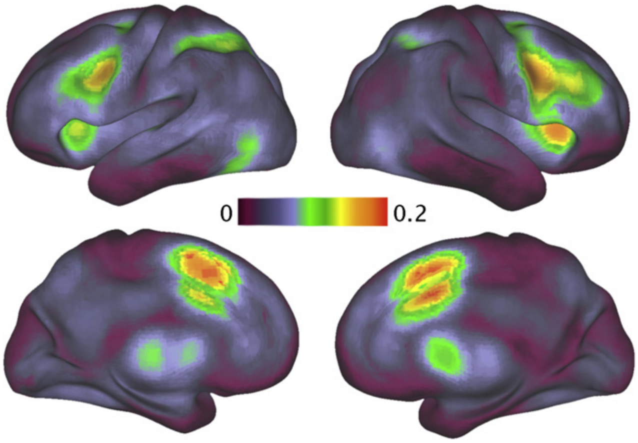

Figure 1. Base Rates for Activation
A rendering of base rates of activation across 3,489 studies in the literature; increasingly hot colors (from yellow to red) reflect more frequent activation across all studies, with the reddest regions active in more than 20% of all studies. Regions of most frequent activation included the anterior cingulate cortex, anterior insula, and dorsolateral prefrontal cortex. Reprinted with permission from Yarkoni et al. (2011).

many vagaries of how researchers use language to describe the mental concepts they are studying; classification will be more accurate for terms that are used more consistently and precisely in the literature. Despite these limitations, the meta-analytic approach has the potential to provide useful insights into the potential strength of reverse inferences.

Decoding Mental States: Toward Formal Reverse Inference

Whereas the kind of reverse inference described above is informal, in the sense that it is based on the researcher's knowledge of associations between activation and mental functions, a more recent approach provides the ability to formally test the ability to infer mental states from neuroimaging data. Known variously as multivoxel pattern analysis (MVPA), multivariate decoding, or pattern-information analysis, this approach uses tools from the field of machine learning to create statistical machines that can accurately decode the mental state that is represented by a particular imaging data set. In the last 10 years, this approach has become very popular in the fMRI literature; for example, in the first 8 months of 2011 there have been more than 50 publications using these methods, versus 41 for the entire period before 2009.

A pioneering example of this approach was the study by Haxby et al. (2001), which showed that it was possible to accurately classify which one of several classes of objects a subject was viewing by using a nearest-neighbor approach, in which a test data set was compared to training data sets obtained for each of the classes of interest. Whereas early work using MVPA focused largely on the decoding of visual stimulus features, such as object identity (Haxby et al., 2001) or simple visual features (Haynes and Rees, 2005; Kamitani and Tong, 2005), it is now clear that more complex mental states can also be decoded from fMRI data. For example, several studies have shown that future intentions to perform particular tasks can be decoded with reasonable accuracy (Gilbert, 2011; Haynes et al., 2007). These studies show that it is possible to quantitatively estimate the degree to which a pattern of brain activation is predictive of the engagement of a specific mental process and to thus provide a formal means to implement reverse inference. They have also provided evidence that activation in some regions may be less diagnostic than is required (and often assumed) for effective reverse inference. For example, neither the ''fusiform face area'' nor the ''parahippocampal place area'' is particularly diagnostic for the stimulus classes that activate them most strongly (faces or scenes respectively) (Hanson and Halchenko, 2008).

Model-Based Approaches

The approach to decoding described above treats the relation between mental states and neuroimaging activation patterns as a data mining problem, estimating relations between the two using statistical brute force. An alternative and more principled approach has been developed more recently, in which the decoding of brain activation patterns is guided by computational models of the putative processes that underlie the psychological function. In one landmark study, Mitchell et al. (2008) showed that it was possible to use the activation patterns from one set of concrete nouns to predict the patterns of activation in another set of untrained words. These predictions were derived using a model that identified semantic features based on correlations between noun and verb usage in a very large corpus of text. By using ''semantic feature maps'' that reflect the activation associated with a semantic feature (which is derived from the mapping of nouns to verbs in the training corpus) predicted activation maps were then obtained by projecting the untrained words into the semantic feature space. These predicted maps were highly accurate, allowing above-chance classification of pairs of untrained words in all of the nine participants.

Another study published by Kay et al. (2008) examined the ability to classify natural images based on fMRI data from the visual cortices. This study estimated a receptive field model for each voxel (based on Gabor wavelets), which modeled the voxel's response along spatial location, spatial frequency, and orientation dimensions, using fMRI data collected while viewing a set of 1,750 natural images. They then applied the model to a set of 120 images that were not included in the training set and attempted to identify which image was being viewed based on the predicted brain activity derived from the receptive field model. The model was highly accurate at decoding which image was being viewed, even when the set of possible images was as large as 1,000. These studies highlight the utility of using intermediate models of the stimulus space to constrain decoding attempts.

In the former cases, the decoding problem was relatively constrained by the presence of a set of test items to be compared, which varied from 2 in the Mitchell et al. (2008) study to up to 1,000 in the Kay et al. (2008). However, subsequent work has shown that it is possible to provide realistic reconstruction of entire images from fMRI data using Bayesian inference with natural image priors, in effect reading the image from the subject's mind. Naselaris et al. (2009) used a model similar to the one described for the Kay et al. (2008) study to attempt to reconstruct images from brain activation. They found that the reconstructions provided by the basic model were not better than chance with regard to their accuracy. However, by using a database of six million randomly selected natural images as priors, it was possible to create image reconstructions that had structural accuracy substantially better than chance. Furthermore, using a hybrid model that also included semantic labels for the images, the reconstructions also had a high degree of semantic accuracy. Another study by Pereira et al. (2011) used a similar approach to generate concrete words from brain activation, using a "topic model" trained on corpus of text from Wikipedia. These studies highlight the utility of model-based decoding, which provides much more powerful decoding abilities via the use of computational models that better characterize mental processes along with statistical information mined from large online databases.

Toward Large-Scale Decoding of Mental States

The foregoing examples of successful decoding are impressive, but each is focused on decoding between different stimuli (images or concrete words) for which the relevant representations are located within a circumscribed set of brain areas at a relatively small spatial scale (e.g., cortical columns). In these cases, decoding likely relies upon the relative activity of specific subpopulations of neurons within those relevant cortical regions or the fine-grained vascular architecture in those regions (see Kriegeskorte et al., 2010 for further discussion of this issue). In many cases, however, the goal of reverse inference is to identify what mental processes are engaged against a much larger set of possibilities. We refer to this here as "large-scale" decoding, in which "scale" refers to both the spatial scale of the relevant neural systems and the breadth of the possible mental states being decoded. Such large-scale decoding is challenging because it requires training data acquired across a much larger set of possible mental states. At the same time, it is more likely to rely upon distributions of activation across many regions across the brain and thus has a greater likelihood of generalizing across individuals compared to the decoding of specific stimuli, which is more likely to rely upon idiosyncratic features of individual brains. Although most previous decoding studies have examined generalization within the same individuals, a number of previous studies has shown that it is possible to generalize across individuals (Davatzikos et al., 2005; Mourão-Miranda et al., 2005; Shinkareva et al., 2008).

In an attempt to test the large-scale decoding concept, we (Poldrack et al., 2009) examined the ability to classify which of eight different mental tasks an individual was engaged in, using statistical summaries of activation for each task compared to rest from each subject. The classifier was tested on individuals who were not included in the training set; the results showed that highly accurate classification was possible, even when generalizing across individuals. Accurate classification was possible using small regions of interest but was greatest using whole-brain data, suggesting that decoding of tasks relied upon both local and global information. Although this work provides a proof of concept for large-scale decoding, true large-scale decoding is still far away; the eight mental tasks tested in this study are but a drop in the very large bucket of possible psychological functions, and each function would probably need to be tested using multiple tasks to ensure independence from specific task features.

A major challenge for large-scale decoding is the lack of a sufficient database of raw fMRI data on which to train classifiers across a large number of different tasks and stimuli. The development of large databases of task-based fMRI data, such as the OpenFMRI project (http://www.openfmri.org), should help provide the data needed for such large-scale decoding analyses. In addition to the need for larger databases, there is also an urgent need for more detailed metadata describing the tasks and processes associated with each data set. The Cognitive Atlas project (http://www.cognitiveatlas.org; Poldrack et al., 2011) is currently developing an ontology that will serve as a framework for detailed annotation of neuroimaging databases, but this is a major undertaking that will require substantial work by the community before it is completed. Until these resources are well developed, the ability to classify mental states on a larger scale is largely theoretical.

Limits on Decoding

Despite the incredible power of these methods to decode mental states from neuroimaging data, some important limits remain. Foremost, decoding methods cannot overcome the fact that neuroimaging data are inherently correlational (cf. Poldrack, 2000), and thus demonstration of significant decoding does not prove that a region is necessary for the mental function being decoded. Lesion studies and manipulations of brain function using methods such as transcranial magnetic stimulation will remain essential for identifying which regions are necessary and which are epiphenomenal. Conversely, a region could be important for a function even if it is not diagnostic of that function in a decoding analysis. For example, it is known that the left anterior insula is critical for speech articulation (Dronkers, 1996). However, given the high base rate of activation in this region (see Figure 1), it is unlikely that large-scale decoding analyses would find this region to be diagnostic of articulation as opposed to the many other mental functions that seem to activate it.

Another important feature of most decoding methods is that they are highly opportunistic, i.e., they will take advantage of any information present that is correlated with the processes of interest. For example, in a recent comparison of univariate and multivariate analysis methods in a decision-making task (Jimura and Poldrack, 2011), we found that many regions showed decoding sensitivity using multivariate methods that did not show differences in activation using univariate methods. This included regions such as the motor cortex, which presumably carries information about the motor response that the subject made (in this case, pressing one of four different buttons). If one simply wishes to accurately decode behavior, then this is interesting and useful, but from the standpoint of understanding the neural architecture of decision making, it is likely a red herring. More generally, it is important to distinguish between predictive power and neurobiological reality. One common strategy is to enter a large number of voxels into a decoding analysis and then examine the importance of each voxel for decoding (e.g., by using the weights obtained from a regularized linear model, as in Cohen et al., 2010). This can provide some useful insight into how the decoding model obtained its accuracy, but it does not necessarily imply that the pattern of weights is reflective of the neural coding of information. Rather, it more likely reflects the match between the coding of information as reflected in fMRI (which includes a contribution from the specific vascular architecture of the region) and the specific characteristics of the statistical machine being used. For example, analyses obtained using methods that employ sparseness penalties (e.g., Carroll et al., 2009) will result in a smaller number of features that support decoding compared to a method using other forms of penalties, but such differences would be reflective of the statistical tool rather than the brain.

Finally, the ability to accurately decode mental states or functions is fundamentally limited by the accuracy of the ontology that describes those mental entities. In many cases of fine-grained decoding (e.g., ‘‘Is the subject viewing a cat or a horse?’’), the organization of those mental states is relatively well defined. However, for decoding of higher-level mental functions (e.g., ‘‘Is the subject engaging working memory?’’), there is often much less agreement over the nature or even the existence of those functions. We (Lenartowicz et al., 2010) have proposed that one might actually use classification to test claims about the underlying mental ontology; that is, if a set of mental concepts cannot be distinguished from one another based on neuroimaging data that are meant to manipulate each one separately, then that suggests that the concepts may not actually be distinct. This might simply reflect terminological differences (e.g., the interchangeable use of ‘‘executive control’’ and ‘‘cognitive control’’) but could also reflect more fundamental problems with theoretical distinctions that are made in the literature.

Whither Reverse Inference?

Given the youth of cognitive neuroscience and the enormity of the problem that we aim to solve, we should use every possible strategy at our disposal, so long as it is valid. Viewed as a means to generate novel hypotheses, I think that reverse inference can be a very useful strategy, especially if it is based on real data (such as the meta-analytic maps from Yarkoni et al., 2011) rather than on an informal reading of the literature. In fact, reverse inference in this sense is an example of ‘‘abductive inference’’ (Pierce, 1998) or ‘‘reasoning to the best explanation,’’ which is widely appreciated as a useful means of scientific reasoning. The problem with this kind of reasoning arises when such hypotheses become reified as facts, as was well stated by the psychologist Daniel Kahneman (Kahneman, 2009):

> The more difficult test, for a general psychologist, is to remember that the new idea is still a hypothesis which has passed only a rather low standard of proof. I know the test is difficult, because I fail it: I believe the interpretation, and do not label it with an asterisk when I think about it. (p. 524)

I would argue that this test is often difficult not just for general psychologists, but also for neuroimaging researchers, who far too often drop the asterisk that should adorn a hypothesis derived from reverse inference until it has been directly tested in further studies.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health grant RO1MH082795 and the National Science Foundation grant OCI-1131441. Thanks to Tyler Davis, Steve Hanson, Rajeev Raizada, Rebecca Saxe, Tom Schonberg, Gaël Varoquax, Corey White, and Tal Yarkoni for helpful comments on a draft of this paper.

REFERENCES

Aguirre, G.K. (2003). Functional imaging in behavioral neurology and cognitive neuropsychology. In Behavioral Neurology and Cognitive Neuropsychology, T.E. Feinberg and M.J. Farah, eds. (New York: McGraw-Hill), pp. 85–96.

Ariely, D., and Berns, G.S. (2010). Neuromarketing: the hope and hype of neuroimaging in business. Nat. Rev. Neurosci. *11*, 284–292.

Carroll, M.K., Cecchi, G.A., Rish, I., Garg, R., and Rao, A.R. (2009). Prediction and interpretation of distributed neural activity with sparse models. Neuroimage *44*, 112–122.

Cohen, J.R., Asarnow, R.F., Sabb, F.W., Bilder, R.M., Bookheimer, S.Y., Knowlton, B.J., and Poldrack, R.A. (2010). Decoding developmental differences and individual variability in response inhibition through predictive analyses across individuals. Front Hum Neurosci *4*, 47.

Davatzikos, C., Ruparel, K., Fan, Y., Shen, D.G., Acharyya, M., Loughead, J.W., Gur, R.C., and Langleben, D.D. (2005). Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. Neuroimage *28*, 663–668.

Dronkers, N.F. (1996). A new brain region for coordinating speech articulation. Nature *384*, 159–161.

Gilbert, S.J. (2011). Decoding the content of delayed intentions. J. Neurosci. *31*, 2888–2894.

Hanson, S.J., and Halchenko, Y.O. (2008). Brain reading using full brain support vector machines for object recognition: there is no ‘‘face’’ identification area. Neural Comput. *20*, 486–503.

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*, 2425–2430.

Haynes, J.-D., and Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat. Neurosci. *8*, 686–691.

Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C., and Passingham, R.E. (2007). Reading hidden intentions in the human brain. Curr. Biol. *17*, 323–328.

Henson, R. (2005). What can functional neuroimaging tell the experimental psychologist? Q. J. Exp. Psychol. A *58*, 193–233.

Iacoboni, M., Freedman, J., Kaplan, J., Jamieson, K.H., Freedman, T., Knapp, B., and Fitzgerald, K. (2007). This is your brain on politics. The New York Times, November 11, 2007. http://www.nytimes.com/2007/11/11/opinion/11freedman.html?pagewanted=all.

Jimura, K., and Poldrack, R.A. (2011). Do univariate and multivariate analyses tell the same story? Neuropsychologia, in press.

Kahneman, D. (2009). Remarks on neuroeconomics. In Neuroeconomics: Decision Making and the Brain, P.W. Glimcher, C.F. Camerer, E. Fehr, and R.A. Poldrack, eds. (London: Academic Press), pp. 523–525.

Kamitani, Y., and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. Nat. Neurosci. *8*, 679–685.

Kay, K.N., Naselaris, T., Prenger, R.J., and Gallant, J.L. (2008). Identifying natural images from human brain activity. Nature *452*, 352–355.

Kriegeskorte, N., Cusack, R., and Bandettini, P. (2010). How does an fMRI voxel sample the neuronal activity pattern: compact-kernel or complex spatiotemporal filter? Neuroimage *49*, 1965–1976.

Lenartowicz, A., Kalar, D., Congdon, E., and Poldrack, R.A. (2010). Towards an ontology of cognitive control. Topics in Cognitive Science *2*, 678–692.

Lindstrom, M. (2011). You love your iPhone. Literally. The New York Times, October 1, 2011, A21. http://www.nytimes.com/2011/10/01/opinion/you-love-your-iphone-literally.html?_r=1.

Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.-M., Malave, V.L., Mason, R.A., and Just, M.A. (2008). Predicting human brain activity associated with the meanings of nouns. Science *320*, 1191–1195.

Mourão-Miranda, J., Bokde, A.L.W., Born, C., Hampel, H., and Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. Neuroimage *28*, 980–995.

Naselaris, T., Prenger, R.J., Kay, K.N., Oliver, M., and Gallant, J.L. (2009). Bayesian reconstruction of natural images from human brain activity. Neuron *63*, 902–915.

Pereira, F., Detre, G., and Botvinick, M. (2011). Generating text from functional brain images. Front Hum Neurosci *5*, 72.

Petersen, S.E., Fox, P.T., Posner, M.I., Mintun, M., and Raichle, M.E. (1988). Positron emission tomographic studies of the cortical anatomy of single-word processing. Nature *331*, 585–589.

Pierce, C.S. (1998). Lectures on pragmatism. In The Essential Pierce, Volume 2: Selected Philosophical Writings, 1893–1913, N. Houser, J.R. Eller, A.C. Lewis, A. De Tienne, C.L. Clark, and D.B. Davis, eds. (Bloomington, IN: Indiana University Press), pp. 331–433.

Poldrack, R.A. (2000). Imaging brain plasticity: conceptual and methodological issues—a theoretical review. Neuroimage *12*, 1–13.

Poldrack, R.A. (2006). Can cognitive processes be inferred from neuroimaging data? Trends Cogn. Sci. (Regul. Ed.) *10*, 59–63.

Poldrack, R.A., Halchenko, Y.O., and Hanson, S.J. (2009). Decoding the large-scale structure of brain function by classifying mental states across individuals. Psychol Sci. *20*, 1364–1372.

Poldrack, R.A., Kittur, A., Kalar, D., Miller, E., Seppa, C., Gil, Y., Parker, D.S., Sabb, F.W., and Bilder, R.M. (2011). The cognitive atlas: toward a knowledge foundation for cognitive neuroscience. Front Neuroinform *5*, 17.

Posner, M.I., Petersen, S.E., Fox, P.T., and Raichle, M.E. (1988). Localization of cognitive operations in the human brain. Science *240*, 1627–1631.

Sarter, M., Berntson, G.G., and Cacioppo, J.T. (1996). Brain imaging and cognitive neuroscience. Toward strong inference in attributing function to structure. Am. Psychol. *51*, 13–21.

Shinkareva, S.V., Mason, R.A., Malave, V.L., Wang, W., Mitchell, T.M., and Just, M.A. (2008). Using FMRI brain activation to identify cognitive states associated with perception of tools and dwellings. PLoS ONE *3*, e1394.

Takahashi, H., Kato, M., Matsuura, M., Mobbs, D., Suhara, T., and Okubo, Y. (2009). When your gain is my pain and your pain is my gain: neural correlates of envy and schadenfreude. Science *323*, 937–939.

Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., and Wager, T.D. (2011). Large-scale automated synthesis of human functional neuroimaging data. Nat. Methods *8*, 665–670.

Young, L., and Saxe, R. (2009). An FMRI investigation of spontaneous mental state inference for moral judgment. J. Cogn. Neurosci. *21*, 1396–1405.