The role of fMRI in Cognitive Neuroscience: where do we stand?
Russell A Poldrack

Functional magnetic resonance imaging (fMRI) has quickly become the most prominent tool in cognitive neuroscience. In this article, I outline some of the limits on the kinds of inferences that can be supported by fMRI, focusing particularly on reverse inference, in which the engagement of specific mental processes is inferred from patterns of brain activation. Although this form of inference is weak, newly developed methods from the field of machine learning offer the potential to formalize and strengthen reverse inferences. I conclude by discussing the increasing presence of fMRI results in the popular media and the ethical implications of the increasing predictive power of fMRI.

Address
UCLA Department of Psychology and Department of Psychiatry and Biobehavioral Sciences, Franz Hall, Box 951563, Los Angeles, CA 90095-1563, United States

Corresponding author: Poldrack, Russell A (poldrack@ucla.edu)

Introduction

fMRI has enjoyed an astounding rise in its use as a tool for cognitive neuroscience research. Since its invention in the early 1990s to the end of 2007, more than 12 000 articles have been published that mention fMRI in the abstract or title (according to PubMed), and this number is now growing by roughly 30–40 papers every week. Many millions of research dollars are being invested in research that uses fMRI and it has made its way into the public eye via high-profile articles in major newspapers. In this article I will review recent work on the nature of inferences that can be supported by neuroimaging data, with a focus on how techniques from the field of machine learning may provide support for a new class of inferences. I will then discuss the ethical implications that have arisen from recent media coverage of fMRI research, specifically with regard to detection of mental states such as lying or political attitudes.

What can we infer from neuroimaging data?

Most neuroimaging research to date has used an approach that Henson [1] has called 'forward inference.' In this approach, conditions that differ in the engagement of some putative mental process are compared and regions that show differences in activation between those conditions are inferred to take part in that mental process. This approach has been remarkably successful, though potential problems with the approach are well known (e.g. [2•,3]). In particular, because it is a correlational approach, one cannot infer that the activated regions are necessary or sufficient for the engagement of the mental process. Indeed, there are well-known examples of cases in which regions that are activated during a task are not necessary for the task. For example, the hippocampus is activated during delay classical conditioning [4], but lesions to the hippocampus do not impair this function [5].

Demonstration of necessity relies upon a manipulation of the region in question, which cannot be achieved with neuroimaging alone. Techniques that allow examination of the effects of manipulating brain function, either indirectly through lesion studies or directly via transcranial magnetic stimulation (TMS), will thus remain a crucial complement to neuroimaging in cognitive neuroscience. It is important to note, however, that inferences from lesion studies are limited by the fact that the brain may often have multiple ways to perform a cognitive process (referred to by Price and Friston [6] as 'degeneracy'). Neuroimaging of lesion patients can help better understand both the ways in which alternate networks may take over task performance [6] as well as how lesions in one region can affect function in other regions [7].

Cognitive neuroscientists have generally adopted a strongly modular approach to structure–function relationships, perhaps driven by the facile leap to localizationist conclusions from lesion and neuroimaging results. Despite the longstanding appreciation for the importance of functional integration within the neuroimaging literature [8,9], the widespread use of functional and effective connectivity analyses has not yet come about. Given that many cognitive processes may be distinguished not by activity in specific regions but by patterns of activity across regions, there is reason for caution regarding many of the inferences that have been driven by highly modular approaches.

Reverse inference

It has become increasingly common to use neuroimaging data to infer the presence of specific mental processes, an approach known as 'reverse inference' [10,11•]. This approach has been particularly common in newer literatures such as neuroeconomics and social cognitive neuroscience, where the fundamental processes underlying

task performance are often unknown. For example, neuroimaging work on moral reasoning has used activation in a set of regions previously associated with emotion to provide evidence for the hypothesis that emotion plays an important role in some kinds of moral judgments [12].

In terms of deductive logic, reverse inference reflects the logical fallacy of affirming the consequent [10,11•]. Such claims are only deductively true if and only if the specific mental process results in the activation in the region of interest, but brain regions observed with fMRI are rarely activated by only one mental process. However, given that the goal of cognitive neuroscientists is to build explanations rather than deductive laws, reverse inferences may provide some useful and important information even if they are not deductively valid. Poldrack [11•] showed that the amount of information provided by reverse inference could be estimated using Bayes' theorem with the BrainMap database [13]. This analysis demonstrated that reverse inference could provide some information about the engagement of specific mental processes, though it is relatively weak because of the fact that activation is rarely selective; that is, regions are often activated by a wide range of mental tasks. This suggests that reverse inference will be most useful when it is used to drive subsequent behavioral or neuroimaging studies, rather than as a direct means to interpreting neuroimaging results.

One of the fundamental questions for cognitive neuroscience is how mental processes are mapped to the brain, which requires some knowledge of what mental processes exist; this is more formally known as an 'ontology' [14]. To date most research has used concepts from cognitive psychology to map onto the brain, but this research has shown that these concepts do not map in a one-to-one fashion to brain regions. One potential contributor to this outcome is that there is no faithful one-to-one mapping of mental processes to specific brain regions; for example, as noted above, specific mental processes may only emerge from the interactions of multiple brain regions [9]. Another potential factor is that there is a faithful mapping of mental processes onto specific brain regions, but that our current ontology for mental processes is incorrect [11•,15•]. Given that much of our current mental ontology has not changed since the 19th century, it would not be surprising if it were scientifically invalid (cf. the fields of physics, chemistry, and cell biology). Answering this question will first require a formal explication of the various versions of the mental ontology; such formalizations would then allow the use of powerful informatics techniques to determine which approaches best fit the data. Recent work has begun to formalize and characterize the structure of cognitive processes using literature mining [16] and projects such as the Cognitive Atlas (http://www.cognitiveatlas.org) aim to use web-based collaboration to better specify the field's current conceptual landscape.

Formalizing reverse inference using multivariate pattern analysis

The reverse inference approach described above is an informal approach to predicting mental states from neuroimaging data. However, the question of how accurately mental states can be predicted from neuroimaging data has been increasingly addressed using pattern classification methods from the fields of statistics and machine learning [17,18,19••]. Whereas standard statistical approaches examine the fit of a model to a sample of data, these pattern classification methods focus on the accuracy of predictions to data that were not used to estimate the model. This can be achieved using 'cross-validation', in which the model is fit to subsets of the data and then tested on the remainder.

The widespread use of pattern classification methods first occurred in the literature on visual object recognition. In response to localizationist claims regarding the representation of specific object categories (e.g. faces or houses) in the ventral temporal cortex, Haxby et al. [20] used a simple pattern classification approach to show that the class of object being perceived could be predicted from patterns of activity in the ventral temporal cortex, even when the regions most selectively activated by a specific class of objects were excluded from the analysis. Another well-known study showed that it was possible to predict the orientation of a visual stimulus based on patterns of activity in visual cortex [21]. More recent work has shown that visual cortical regions contain information sufficient to identify specific visual scenes [22] and specific faces [23]. Whereas much of the work in this area has regarded visual information processes, other recent work has shown that such approaches can also be used to detect high-level cognitive processes such as intention [24•], deception [25], and word meaning [26••]. Similar methods have also been applied to the decoding of neural signals in other neuroscientific domains, such as in the context of invasive recordings in animals [27] and in the context of noninvasive brain–computer interfaces using EEG (e.g. [28]).

The advent of pattern classification methods for fMRI analysis promises to change the focus of neuroimaging studies from the detection of activation to the quantification of information that is present in the neuroimaging signal, and from a focus on specific regions to large-scale networks [29••]. Every neuroimaging researcher knows that the activation patterns associated with specific tasks are rarely diagnostic (i.e. specifically predictive) of that task. With pattern classification methods it is possible to identify patterns of activity that are specifically diagnostic of a particular task or stimulus type [30]. Such an approach has the potential to provide much greater specificity to neuroimaging results.

Although most work to date has focused on classification within individuals, some studies have shown that it can be

possible to successfully classify mental states across individuals [25,31]. The ability to classify across individuals would provide a basis for formal reverse inference, but this would require a database of activation patterns against which any particular dataset could be compared. Although such classification has been demonstrated using a handful of tasks (Poldrack, Halchenko, and Hanson, unpublished data), it is not known how well it would scale to a large number of potential mental states. In addition, such large-scale classification would require databases that contain whole-brain activation patterns across a wide range of mental states. The only current database containing a large enough number of studies to be useful is the BrainMap database, but the representation in this database is too impoverished to support pattern classification (i.e. it stores only the location of reported peak activations, and does not store patterns for individual subjects).

Neuroimaging, ethics, and the media

The ability to use fMRI to detect mental states raises important ethical questions, which have come to the fore in the context of lie detection. A number of studies have examined the ability to detect particular forms of instructed lying using fMRI, with several studies demonstrating the ability to accurately detect lying across individuals [25,32]. On the basis of these results, at least two companies have been formed to sell fMRI lie detection services. This research has been systematically reviewed by Greely and Illes [33••] and Sip et al. [34•], both of whom conclude that the results are far from providing support for the kinds of claims that the proponents of fMRI lie detection wish to make. Greely and Illes also lay out a roadmap for the regulation of commercial uses of fMRI for lie detection.

There are also more subtle but equally important ethical questions related to the presentation of fMRI data in the media. It seems difficult today to open a newspaper without reading a story about the latest finding using brain imaging. The ability to see inside the working human mind has captured the popular imagination and the press has jumped on these results with great vigor, if not great care [35]. In some cases, the normal methods of peer review have been subverted in the name of publicity. In one example, a group including neuroscientists, political scientists, and market researchers published an op-Ed in the New York Times [36], which presented novel fMRI research that examined the response of uncommitted voters to videos of US presidential candidates. The conclusions relied largely upon informal reverse inference; for example, ''When we showed subjects the words 'Democrat,' 'Republican' and 'independent,' they exhibited high levels of activity in the part of the brain called the amygdala, indicating anxiety''. This article provoked a substantial response from neuroscientists [37,38] and others [39,40], expressing concern that the standard peer review process for scientific publication

was subverted as well as criticism of the unchecked use of reverse inference.

One particular problem with the popularization of neuroimaging data is that these data seem to have a disproportionately strong persuasive impact [41]. Recent psychological studies have shown that the presentation of a brain image can increase the reader's judgment of the quality of the reasoning [42••] of the article, and even verbal descriptions of neuroimaging results can make weak arguments more persuasive even when they are irrelevant to the argument [43•]. These results place added responsibility on neuroimaging researchers to present their work responsibly in the press.

Conclusions

As fMRI has matured as an imaging technology and the body of existing research has grown, it has become increasingly possible to use fMRI data to 'read' mental states from brain activity, first informally and increasingly using formal methods from machine learning. I believe that these methods will provide the basis for the next generation of neuroimaging in combination with more detailed models of neural connectivity and computational modeling. There is concern, however, that a failure to appreciate and directly address the ethical implications of this work could lead to a backlash, including regulations that could hobble fMRI research.

Acknowledgements

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

• of special interest
•• of outstanding interest

1. Henson R: **Forward inference using functional neuroimaging: dissociations versus associations**. *Trends Cogn Sci* 2006, **10**:64-69.

2. Henson R: **What can functional neuroimaging tell the**
•   **experimental psychologist?** *Q J Exp Psychol A* 2005, **58**:193-233.
This paper provides the most thorough and detailed explication to date of the ways in which neuroimaging can be used to make inferences about psychological processes.

3. Poldrack RA: **Imaging brain plasticity: conceptual and methodological issues — a theoretical review**. *Neuroimage* 2000, **12**:1-13.

4. Knight DC, Smith CN, Cheng DT, Stein EA, Helmstetter FJ: **Amygdala and hippocampal activity during acquisition and extinction of human fear conditioning**. *Cogn Affect Behav Neurosci* 2004, **4**:317-325.

5. Gabrieli JDE, Carrillo MC, Cermak LS, Mcglinchey-Berroth R, Gluck MA, Disterhoft JF: **Intact delay–eyeblink classical conditioning in amnesia**. *Behav Neurosci* 1995, **109**:819-827.

6. Price CJ, Friston KJ: **Scanning patients with tasks they can perform**. *Hum Brain Mapp* 1999, **8**:102-108.

7. Alho K, Woods DL, Algazi A, Knight RT, Naatanen R: **Lesions of frontal cortex diminish the auditory mismatch negativity**. *Electroencephalogr Clin Neurophysiol* 1994, **91**:353-362.

8. Friston KJ: **Functional and effective connectivity in neuroimaging: a synthesis**. *Hum Brain Mapp* 1994, **2**:56-78.

9. McIntosh AR: **Towards a network theory of cognition**. *Neural Netw* 2000, **13**:861-870.

10. Aguirre GK: **Functional imaging in behavioral neurology and cognitive neuropsychology**. In *Behavioral Neurology and Cognitive Neuropsychology*. Edited by Feinberg TE, Farah MJ. McGraw-Hill; 2003.

11. Poldrack RA: **Can cognitive processes be inferred from**
• **neuroimaging data?** *Trends Cogn Sci* 2006, **10**:59-63.
This was the first paper to systematically examine the support that is provided by reverse inference. It presents an analysis of the BrainMap neuroimaging database, which demonstrated that reverse inferences of the kind commonly used in the literature provide relatively weak evidence regarding the engagement of specific mental processes.

12. Greene JD, Sommerville RB, Nystrom LE, Darley JM, Cohen JD: **An fMRI investigation of emotional engagement in moral judgment**. *Science* 2001, **293**:2105-2108.

13. Laird AR, Lancaster JL, Fox PT: **BrainMap: the social evolution of a human brain mapping database**. *Neuroinformatics* 2005, **3**:65-78.

14. Bard JB, Rhee SY: **Ontologies in biology: design, applications and future challenges**. *Nat Rev Genet* 2004, **5**:213-222.

15. Price CJ, Friston KJ: **Functional ontologies for cognition: the**
• **systematic definition of structure and function**. *Cogn Neuropsychol* 2005, **22**:262-275.
This was the first paper to argue for the utility of formal ontologies of cognitive processes. It demonstrates how such ontologies could be used to more systematically map mental function onto neural structure.

16. Sabb FW, Bearden CE, Glahn DC, Parker DS, Freimer N, Bilder RM: **A collaborative knowledge base for cognitive phenomics**. *Mol Psychiatry* 2008, **13**:350-360.

17. Haynes JD, Rees G: **Decoding mental states from brain activity in humans**. *Nat Rev Neurosci* 2006, **7**:523-534.

18. Norman KA, Polyn SM, Detre GJ, Haxby JV: **Beyond mind-reading: multi-voxel pattern analysis of fMRI data**. *Trends Cogn Sci* 2006, **10**:424-430.

19. O'Toole AJ, Jiang F, Abdi H, Penard N, Dunlop JP, Parent MA:
•• **Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data**. *J Cogn Neurosci* 2007, **19**:1735-1752.
This review paper provides an outstanding overview of the conceptual and methodological issues involved in multivoxel pattern analysis of fMRI data. It provides a measured survey of both the potential power of these methods and the caveats that must be kept in mind in evaluating this research.

20. Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P: **Distributed and overlapping representations of faces and objects in ventral temporal cortex**. *Science* 2001, **293**:2425-2430.

21. Kamitani Y, Tong F: **Decoding the visual and subjective contents of the human brain**. *Nat Neurosci* 2005, **8**:679-685.

22. Kay KN, Naselaris T, Prenger RJ, Gallant JL: **Identifying natural images from human brain activity**. *Nature* 2008, **452**:352-355.

23. Kriegeskorte N, Formisano E, Sorger B, Goebel R: **Individual faces elicit distinct response patterns in human anterior temporal cortex**. *Proc Natl Acad Sci U S A* 2007, **104**:20600-20605.

24. Haynes JD, Sakai K, Rees G, Gilbert S, Frith C, Passingham RE:
• **Reading hidden intentions in the human brain**. *Curr Biol* 2007, **17**:323-328.
In this paper, the authors used pattern classification to predict complex cognitive states. Subjects decided whether to add or subtract two numbers; analyses showed that these decisions could be predicted with about 70% accuracy using activity from regions in the medial prefrontal cortex. The results show that even high-level cognitive states can be predicted from fMRI data.

25. Davatzikos C, Ruparel K, Fan Y, Shen DG, Acharyya M, Loughead JW, Gur RC, Langleben DD: **Classifying spatial patterns of brain activity with machine learning methods: application to lie detection**. *Neuroimage* 2005, **28**:663-668.

26. Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL,
•• Mason RA, Just MA: **Predicting human brain activity associated with the meanings of nouns**. *Science* 2008, **320**:1191-1195.
This remarkable paper shows how fMRI activation patterns for new words can be predicted based on fMRI activation to a small set of words along with the co-occurrence patterns of words in a large text corpus. It proposes that the representation of words is composed from a 'basis set' of semantic features that have different neural signatures, which can be detected using fMRI. The results demonstrate the power of classifier methods to characterize the predictive relation between stimuli and brain activation patterns.

27. Lin L, Osan R, Tsien JZ: **Organizing principles of real-time memory encoding: neural clique assemblies and universal neural codes**. *Trends Neurosci* 2006, **29**:48-57.

28. Blankertz B, Dornhege G, Krauledat M, Muller KR, Curio G: **The non-invasive Berlin brain–computer interface: fast acquisition of effective performance in untrained subjects**. *Neuroimage* 2007, **37**:539-550.

29. Kriegeskorte N, Goebel R, Bandettini P: **Information-based**
•• **functional brain mapping**. *Proc Natl Acad Sci U S A* 2006, **103**:3863-3868.
This paper lays out a rationale for 'information-based', as opposed to 'activation-based' neuroimaging. It develops a method for localized estimation of the information that is present in fMRI data, and shows that these methods are more powerful than standard activation mapping techniques for the detection of signals.

30. Hanson SJ, Halchenko YO: **Brain reading using full brain support vector machines for object recognition: there is no "face" identification area**. *Neural Comput* 2008, **20**:486-503.

31. Mourao-Miranda J, Bokde AL, Born C, Hampel H, Stetter M: **Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data**. *Neuroimage* 2005, **28**:980-995.

32. Kozel FA, Johnson KA, Mu Q, Grenesko EL, Laken SJ, George MS: **Detecting deception using functional magnetic resonance imaging**. *Biol Psychiatry* 2005, **58**:605-613.

33. Greely HT, Illes J: **Neuroscience-based lie detection: the urgent**
•• **need for regulation**. *Am J Law Med* 2007, **33**:377-431.
This is the most systematic review to date of the scientific and legal issues surrounding lie detection using neuroimaging methods. The paper reviews the literature on lie detection using fMRI, concluding that the current evidence does not support broad claims about the usefulness of fMRI for reliable lie detection. It also discusses how current law regarding polygraphy relates to fMRI lie detection, and argues cogently for the need for new legislation to address commercial uses of fMRI for lie detection.

34. Sip KE, Roepstorff A, McGregor W, Frith CD: **Detecting deception:**
• **the scope and limits**. *Trends Cogn Sci* 2008, **12**:48-53.
This paper argues that research on lie detection using fMRI has not addressed some of the most important issues related to deception and lying.

35. Racine E, Bar-Ilan O, Illes J: **fMRI in the public eye**. *Nat Rev Neurosci* 2005, **6**:159-164.

36. Iacoboni M, Freedman J, Kaplan J, Jamieson KH, Freedman T, Knapp B, Fitzgerald K: **This is your brain on politics**. *New York Times*, November 11, 2007.

37. Aron A, Badre D, Brett M, Cacioppo J, Chambers C, Cools R, Engel S, D'Esposito M, Frith C, Harmon-Jones E *et al.*: **LETTER: politics and the brain**. In *New York Times*, 2007 (http://www.nytimes.com/2007/11/14/opinion/lweb14brain.html).

38. Farah MJ: **This is your brain on politics?** In *Neuroethics & Law Blog*, 2007 (http://kolber.typepad.com/ethics_law_blog/2007/11/this-is-your-br.html).

39. **Editorial. Mind games**. *Nature* 2007, **450**:457.

40. Engber D: **Neuropundits gone wild!** In *Slate*, 2007 (http://www.slate.com/id/2177885/).

41. Bloom P: **Seduced by the flickering lights of the brain**. In *Seed*, 2006 (http://www.seedmagazine.com/news/2006/06/seduced_by_the_flickering_ligh.php).

42. McCabe DP, Castel AD: **Seeing is believing: the effect of brain**
•• **images on judgments of scientific reasoning**. *Cognition* 2008,
107:343-352.
This study provides a compelling demonstration that brain images have
powerful effects on judgments about scientific evidence. Subjects were
presented with brief articles describing neuroscientific results, which
included either brain images or bar graphs. Presentation of brain images
resulted in an increase in the subjects' willingness to believe the con-
clusions of the article.

43. Weisberg DS, Keil FC, Goodstein J, Rawson E, Gray JR: **The**
• **seductive allure of neuroscience explanations**. *J Cogn*
Neurosci 2008, **20**:470-477.
This study presented subjects with psychological explanations that were
either accompanied or unaccompanied by irrelevant neuroscientific
information. The presence of neuroscientific information caused nonex-
pert subjects to judge the arguments are more satisfying, suggesting that
these kinds of explanations may have particularly powerful persuasive
effects.