Interpreting brain images – reflections on an adolescent field

Foundational Issues in Human Brain Mapping, edited by Stephen José Hanson and Martin Bunz, The MIT Press, 2010. 321pp, ISBN 978-0-262-51394-4.

Nikolaus Kriegeskorte
Medical Research Council, Cognition and Brain Sciences Unit, Cambridge, CB2 7EF, UK
Corresponding author: Kriegeskorte, N. (Nikolaus.Kriegeskorte@mrc-cbu.cam.ac.uk)

Functional brain imaging is maturing, but still adolescent. It has developed a rich toolbox of experimental and data analytical techniques and is addressing an ever expanding range of questions about brain and mind – at various levels of methodological rigor. Some of these questions (e.g. romantic love) are difficult to pin down with science. Occasionally results are naively overinterpreted in scientific papers and in the media. It is appropriate then to reflect on our basic assumptions.

This edited book is a useful collection of conceptual and methodological arguments on how to best use imaging to learn about cognition and brain function. The issues range from experimental design and analysis to the theoretical interpretation of the results, spanning multiple disciplines, including statistics, computational modeling, cognitive and brain theory, and philosophy.

Imaging seems to explain the fluff of the psyche at the level of the hardware. And it combines the prestige of serious science with the broad appeal of intuitive images. This combination is dangerously seductive. The brain blob has the power to make us believe, however tenuous its link to the proposition in question.

But brain images are not like photos: direct and simple reflections of their content matter. We mustn't jump from colored blob to mental conclusion. Instead we need to consider the intervening inferential steps: the blob through the statistics reflects the imaging signal, which reflects the hemodynamic response to neuronal activity, which, in turn, may or may not underlie the mental phenomenon (Roskies; parenthetical names refer to chapter authors). These perils notwithstanding, our intuition is fundamentally correct: Brain images really do afford discovery ('Will any region be found?' 'If so, which one?') and substantial theoretical insight on brain information processing.

Since the cognitive revolution, we have been constructing theories about information processing in the brain. Initially our models of cognition were based on behavioral data alone. Despite ingenious methods of inferring internal processes, cognitive theory is vastly underconstrained by behavioral data: there are many different theories consistent with the data. Brain imaging can help not only to localize functions anatomically, but also to better constrain theories at the cognitive and neural levels (Coltheart; alternative perspectives by Mole and Klein; Harman; Loosemore and Harley; and Bechtel and Richardson).

One challenge of engineering (or reverse engineering) an information processing system is functional decomposition: how is the complex process to be divided into functional subcomponents implemented in separate physical parts of the system?

In building computers and algorithms, we divide the system into modules such that interactions across boundaries are limited. This allows us to reason about the system at a higher level of description, where we can safely disregard the intricate interactions within each module. Are we carving nature at the joints when describing cognition (and the brain) in terms of modules? Or are we just carving it? Brains are built by evolution, development, and learning, processes very different from design by human sequential conscious thought. These processes may not require modularity to the same degree (or at all) to build a well-functioning information processing system.

Undeterred by such theoretical quibbles and emboldened by the specificity of behavioral deficits resulting from localized brain damage, imaging began by attempting to localize the modules of cognitive theory in the brain. Modularity in its strong form proved difficult to demonstrate in general (examples in early sensory and motor cortex notwithstanding). However, the paradigm of finding brain regions with some degree of functional specialization has been very successful overall. The brain is not modular in Fodor's sense,[1] but it is also clearly not equipotential in Lashley's sense.[2]

In the subtraction technique (critically discussed by Poldrack), two cognitive tasks are designed to differ by a single component process. The activity patterns are then subtracted to localize that process. The clean boundary of a blob resulting from statistical thresholding of a subtraction map must not be taken to suggest that we are looking at a sharply defined anatomical structure that corresponds to a functional module. However, a functional decomposition that yields more robust localization results may come closer to capturing the functional organization of the brain.

Once a region's specialization has been tentatively established, we may want to characterize its functional properties further. To this end, we may repeat the initial experiment to localize the region in each subject, and then perform a new experiment to study the region's activity during different cognitive states. On the positive side (Saxe et al.), this functional localizer technique greatly simplifies design, analysis, and interpretation by focusing on a particular hypothesis.[3] On the negative side (Friston et al.), the

functional localizer technique promotes the selective reporting of results based on a potentially premature acceptance of the region as a natural entity with well defined boundaries.[4] The approach often forgoes an exploration of alternative parcellations into functional units, richer analyses of other brain regions, and a systematic investigation of the interactive effect of the factor the region is defined by and the other factors of interest. The technique, thus, promotes a confirmation bias.

More sophisticated factorial designs provide all the functionality of localizers, but additionally allow us to systematically explore the effect of interactions between the region-defining factor and the other factors of interest. However, the more comprehensive exploration of the space of possible effects comes at a loss in statistical power.

While functional localizers promote one kind of selection bias (the selective reporting of correct results or confirmation bias), they help prevent another type of selection bias, which can arise when the same data set is used to define the region and analyze its functional properties in detail (Vul and Kanwisher; with a contrasting perspective from Poldrack and Mumford): Voxels whose noise component is more consistent with the selection criterion are more likely to be selected. This can bias activation estimates and yield spuriously significant, incorrect results.[5]

Perceptual and cognitive content has long been thought to be represented in distributed codes. While the idea of modularity motivates functional localization, the idea of distributed representation motivates the analysis of continuous activity patterns for the information they contain (Haxby).[6] Note that distributed codes could reside within modules, so distributed and localist perspectives, though contrasting, are not contradictory (Bunzl et al.). The regional-average activation typically analyzed in brain imaging merely indicates the "involvement" of a region in some function. Pattern-information analysis allows us to look into each region and characterize its representational content.[7]

The book covers a range of additional topics. Hanson and Glymour explain how to analyze causal influences between brain regions. Biswall reviews the study of slow resting-state activity fluctuations correlated between different parts of the brain. Grill-Spector reviews the study of effects of stimulus change and their interpretation in terms of neuronal tuning (with contrasting remarks by Poldrack). Poline et al. discuss how to address intersubject variability and brain activity as an intermediate endophenotype elucidating the relationship between genes and behavior.

Unlike a textbook, this volume speaks with many voices, highlighting different perspectives by means of dialogue within and across chapters. The book will be of interest to graduate students as well as postdoctoral and senior scientists. The issues are foundational indeed, and deserve our sustained attention.

References

1 Fodor JA (1983) *The Modularity of Mind: An Essay on Faculty Psychology.* Bradford/MIT Press.

2 Lashley KS (1929/1963) *Brain Mechanisms and Intelligence: A Quantitative Study of Injuries to the Brain.* Dover.

3 Kanwisher N, McDermott J, Chun MM (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J Neurosci* 17, 4302-11.

4 Friston KJ, Rotshtein P, Geng JJ, Sterzer P, Henson RN (2006) A critique of functional localisers. *Neuroimage* 30, 1077-87.

5 Kriegeskorte N, Lindquist MA, Nichols TE, Poldrack RA, Vul E (in press) Everything you never wanted to know about circular analysis, but were afraid to ask. *J. Cereb. Blood Flow Metab.* doi:10.1038/jcbfm.2010.86

6 Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293, 2425-30.

7 Mur M, Bandettini P, Kriegeskorte N (2009) Revealing representational content with pattern-information fMRI – an introductory guide. Soc. Cogn. Affect. Neurosci. 4, 101-9.