Cognitive Ontology and Region- versus Network-Oriented Analyses

Colin Klein*

The interpretation of functional imaging experiments is complicated by the pluripotency of brain regions. As there is a many-to-one mapping between cognitive functions and their neural substrates, region-based analyses of imaging data provide only weak support for cognitive theories. Price and Friston argue that we need a 'cognitive ontology' that abstractly categorizes the function of regions. I argue that abstract characterizations are unlikely to be cognitively interesting. I argue instead that we should attribute functions to regions in a context-sensitive manner. I review recent meta-analyses that approach fMRI data in this light and argue that they have revisionary potential.

1. The Problem of Reverse Inference. Different parts of the brain make different contributions to cognition. No part of the brain works alone, however: a cognitive task is always performed by a network of brain regions working in concert. Both of these claims should be unremarkable. When it comes to neuroimaging, though, these truisms have received markedly different emphases. Network-oriented analyses are becoming increasingly common (Pessoa 2008; Ramsey et al. 2009). For roughly the first decade and a half of neuroimaging (NI), however, the focus was on the function of brain regions considered in isolation from the broader neural context.

Region-oriented analyses should be familiar to anyone who has looked at NI studies. Typical studies do one of two things. On the one hand, they localize cognitive processes to particular brain areas: differential regional activity between two cognitive tasks is taken as evidence about the function of that region. So, for example, a region of left posterior lateral fusiform gyrus (PLF) is more active when subjects passively view words than when they view checkerboards. On the basis of this, Cohen et al. (2000) assigned the function of 'visual word form identification' to PLF. On the other hand, ac-

*To contact the author, please write to: 1420 University Hall, MC 267, University of Illinois at Chicago, 601 S. Morgan St., Chicago, IL 60607; e-mail: cvklein@uic.edu.

Philosophy of Science, 79 (December 2012) pp. 952–960. 0031-8248/2012/7905-0005$10.00
Copyright 2012 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/667843 Published online by Cambridge University Press

tivity in a particular area is used as an operationalization of a given cognitive process. Given Cohen et al.'s identification, activity in the PLF can be used as a marker for visual word form identification in future experiments. The point of this operationalization is to test cognitive theories: if we see increased PLF activation in task *A*, we can rule out theories that claim that *A* has nothing to do with word form identification. These complimentary strategies have come to be known respectively as *forward inference* and *reverse inference* (Poldrack 2006).

As our stock of NI experiments has grown, a problem has arisen. Using regional activation *R* as a marker for cognitive function *F* is good practice only if $p(F|R)$ is relatively high. For many brain regions, that does not seem to be the case. As Price and Friston point out, PLF is active in a wide variety of other contrasts, many of which have nothing to do with words: naming pictures versus reading object names, making action decisions versus size decisions on objects, naming color patches, decoding braille words versus nonwords, and so on (2005). Thus, saying that *the* function of PLF is word form identification is hasty: PLF does many things.

Similar examples could be found for nearly any brain region. Given the way we have carved up cognitive functions, brain regions appear to be *pluripotent*: that is, there is many-to-one mapping between functions and brain regions. For any particular function we assign a brain region, activity in that region will be a poor indicator of which cognitive process is engaged: $p(F|R)$ will be generally quite low (Poldrack 2006).

This has become known as the Problem of Reverse Inference. I think the name is a bit unfortunate, however. It suggests that so-called forward inferences, from cognitive theory to brain function, are basically sound and that the problem occurs when we try to move back from brain activity to cognitive function. But arguably, pluripotency is equally troubling for the forward inference step. We know that the PLF does many things; what could possibly justify calling it the visual word form area? One might hope that functional attributions would also take us some way toward understanding the organization of the brain. It is hard to know how NI is supposed to help achieve that goal if we cannot even say confidently what particular brain regions are doing.

2. Price and Friston. The problem of reverse inference might push one toward simple skepticism about NI or at least about the prospect of localizing brain function. In a thought-provoking article, Price and Friston (2005) offer a more optimistic response to the problem.

To begin, note that the function of a thing can be described at varying levels of abstraction. An analogy: many computer operating systems include a subroutine for doing fast Fourier transforms (call it *S*). What is the function of *S*? Well, it depends on how we describe it. We can give many specific

https://doi.org/10.1086/667843 Published online by Cambridge University Press

functional attributions. In image compression programs, the function of *S* is to identify high-frequency components that can be elided without loss of fidelity. In audio programs, the function of *S* is to separate out different audio channels. Each of these attributions is true and useful, but their truth is context sensitive and so holds only for a particular program. However, *S* can be described at a quite general level: every instance of *S* translates input from the time (space) domain to the frequency (phase) domain (or vice versa). At this level of abstraction, *S* does one thing every time it is called. Note that this general functional attribution is still context sensitive (it says something about what *S* contributes to programs) but is true in any context in which *S* is called.

So is *S* pluripotent? That depends on the level of abstraction we have chosen. In terms of specific functions, *S* does many things; in terms of general functions, it does one thing. That suggests a possible solution to the problem of reverse inference. Perhaps brain regions only appear pluripotent because we have not specified their function in suitably general terms. Make it abstract enough, and we will find that brain regions do only one thing after all.

Price and Friston take exactly this line. As they see it, the problem of reverse inference really shows a problem with existing *cognitive ontologies*: that is, in the stock of basic functions and relations that we use to describe cognitive tasks. Our existing ontologies are likely to be deficient in two ways (2005, 268). First, they do not include functions at a suitably abstract level. Given a region, our primary goal should be to attribute it a "function that explains all patterns of activation" (268). Every task that activates PLF seems to require the integration of perceptual and motor information. The function of PLF, then, is 'sensorimotor integration'. This is true whenever PLF is active; at this level of abstraction, PLF is not pluripotent. That in turn gives a nice solution to the problem of reverse inference that I outlined above. When we see PLF activation, we can be sure—now with deductive certainty—that sensorimotor integration is occurring. In general, a set of suitably abstract functional labels will allow two-way prediction: activation in structures will predict function and vice versa (269).

Price and Friston further argue that our cognitive ontologies are deficient because they include a lot of illegitimate specific functional labels (like 'visual word form identification'). These specific functional attributions are likely to be misleading, in part because the specific function only occurs within a larger neural context. So, for example, Price and Friston claim that there is really no area specifically devoted to visual word form processing: word processing arises only "from the interactions among early visual and later reading stages" (2005, 268). By purging these specific functions in favor of general functions, we will put our NI practice on a firm foundation.

https://doi.org/10.1086/667843 Published online by Cambridge University Press

3. Objections to Price and Friston. Price and Friston's picture is undeniably attractive. Science is powerful in part because it abstracts away from details to give general explanations. Closer inspection reveals some serious problems, though.

Consider the putative attribution of 'sensorimotor integration' to PLF. That is undoubtedly true. It is also extremely vague. As Price and Friston note, other parts of the brain also do sensorimotor integration (2005, 267); at some level of abstraction, that is what nearly all of the cortex does. So this level of abstraction does not allow us to answer some crucial questions: for example, why it is PLF we see in reading tasks, rather than some other region. Surely that is also a question we would like to answer. By focusing only on the most abstract level, then, Price and Friston seem to give up on answering a lot of questions that depend on the details.

One might object that this is uncharitable: 'sensorimotor integration' is really a placeholder, to be fleshed out as appropriate by further research. I think the problem is deeper than that. Consider a different analogy. The pistons on many diesel trucks have two specific functions. Most of the time, they compress a fuel-air mixture to the point of detonation and transmit the generated power to the crankshaft. On trucks equipped with engine brakes, the pistons also have a second function: when the engine brake is engaged, the pistons use power from the wheels to compress air in the cylinder, slowing the truck. Which function the piston performs depends on things external to it: whether it is powering or slowing the truck depends on the ignition system and the valve timing.

Both of the context-specific functions of the piston are straightforward, easy to understand, and useful to cite. What is the most general description that covers both cases? Well, it seems to be something like, "The job of the piston is either to speed the truck or to slow it down or to maintain a steady speed." That level of description, I argue, is essentially useless: it is true of everything under the hood. So we have a counterexample to Price and Friston's claim: sometimes abstract functional descriptions are not especially telling.

A similar worry applies to Price and Friston's sensorimotor integration. The most general function that can be attributed to a region is not guaranteed to be cognitively interesting. Specific functional attributions, when available, provide relatively strong constraints on cognitive theories. The more we abstract away from those details, the less constrained our cognitive theorizing becomes. To put it more bluntly: suppose we see PLF activation and so know that there is some sensorimotor integration going on. It is hard to know what cognitive theory could possibly conflict with that: at that level of abstraction, any theory looks like it will be compatible with PLF activation.

https://doi.org/10.1086/667843 Published online by Cambridge University Press

4. Context-Sensitive Reverse Inference. I agree with Price and Friston that we need to rethink our cognitive categories. There is a different lesson we might draw. To begin, note that the second half of Price and Friston's claim—that we should abandon context-specific functional attributions—does not seem to be terribly well motivated. Just because we can give a very abstract functional description does not mean that more specific attributions will not also be true and worth caring about. Because specific attributions are context sensitive, we have to use them with care: we are only justified in appealing to them if we are sure we are in the same context as the original attribution. But if we are confident, we may argue as follows:

1. The function of R in context C is F
2. R is active in C′
3. $C' \subseteq C$

$\therefore$ The function of R in C′ is F.

That is, if we restrict ourselves to the same context, we can infer that a specific cognitive function is employed.

A quick note about premise three: I have put it in terms of subsets rather than identity because contexts are themselves describable in more or less abstract terms. The very same task might be described as "reading the word 'dog,' " "reading an alphabetic word," "reading," and so on. This hierarchical structure is important, and I will return to it.

The argument pattern is valid. Why is reverse inference difficult in practice? Well, in part it is because experimenters have not been terribly careful about specifying the context they are in. Nor have they taken care to determine whether two nominally distinct tasks really count as part of the same context. So in practice, reverse inference is rarely deployed in a convincing way.

It is worth noting that this not just a formal problem but also an underappreciated practical one. In the sorts of experiments I have described, the primary data are statistically significant differences in the blood-oxygen-level-dependent signal between tasks. How we interpret those differences depends on how similar we think the task contexts are. If the two contexts are relevantly distinct, then a difference in brain activation in some region shows something about brain regions that perform wholly distinct functions. If one assumes, however, that different tasks are simply variations on the same theme, then differences in activation indicate only differential computational demands on the same function. As different cognitive theories carve up contexts differently, the very same data can be interpreted in different ways.

Examples abound in the literature. Drawing from behavioral data, Kanwisher and colleagues hypothesized that face processing was performed by a specialized module (Kanwisher 2000). They noted differential activation

https://doi.org/10.1086/667843 Published online by Cambridge University Press

in fusiform face area during the viewing of faces as compared to houses and interpreted this as support for their hypothesis (Kanwisher et al. 1997). Ishai and colleagues, by contrast, hypothesized that face and house recognition were two species of the same distributed recognition process. In an experiment examining the same contrast, they argued that the data supported a single genus of computation, with differences in activation showing something about the computation underlying that process (Ishai et al. 1999; Haxby et al. 2001).

Another example: as Green and others have noted, differential inferior prefrontal activation in second- versus first-language syntactic processing is prima facie consistent with either of two hypotheses: that different languages are differentially represented or else that second-language syntax engages the same circuits as the primary language but less efficiently (Green 2003; Green, Crinion, and Price 2006; Abutalebi 2008). In both cases, the disagreement is not (just) about how to carve up brain functioning. It is a disagreement about just what tasks the subject is being asked to perform. A disagreement at that level trickles down to the interpretation of differential activity, preventing a satisfying resolution.

These controversies suggest a second deeper lesson. The problem with reverse inference is not that people have been sloppy in specifying the contexts in which activation takes place. Rather, it is that we might be profoundly mistaken about which contexts there are.

5. Networks and Contexts. That brings us back to brain networks and the question of region- versus network-analysis styles. I have been deliberately vague about what the 'context' in context-specific attributions refers to. Strictly speaking, it must refer to *neural* context: that is, to the overall network in which a region is participating. (For convenience, I will speak of networksas simply sets of brain regions, but distinct brain networks can also be formed by changes in functional connectivity between the same regions.) In practice, experimental tasks are used as proxies for information about networks: one assumes that similar tasks engage the same network, and distinct tasks engage at least partially distinct networks.

I have argued that using experimental tasks as a proxy for networks is a shaky practice: we should be confident in it only if we think that our preexisting task ontologies actually correspond to real similarities and differences in brain function. That might be false. Further, since there is often reasonable debate about how to distinguish tasks, there is a fortiori reasonable debate about how brain networks might be distinguished.

Although our assumptions about the structure of tasks might be wrong, they are still testable assumptions. That is, we can look to see whether tasks we think are similar (or distinct) activate similar (or distinct) networks. Ev-

https://doi.org/10.1086/667843 Published online by Cambridge University Press

idence for similar activation is evidence that we have divided up tasks correctly; evidence to the contrary might suggest useful revision of how we think about cognitive tasks.

This can be done in a loose and informal way. For example, Klein (2011) argues that when you look at the set of brain regions consistently activated across moral decision-making tasks, you find a relatively stable collection that includes the precuneus and posterior cingulate cortex, the temporoparietal junction, and the ventromedial prefrontal cortex. This network is commonly activated in tasks that require social self-projection: crudely stated as thinking of oneself in another's shoes (Buckner and Carroll 2007). The best explanation of the activation seen in moral decision making, then, is that moral reasoning contexts are specific instances of the more general context of projective social reasoning. That suggests in turn that differential activations between different types of moral dilemma do not represent different brain pathways specialized for different types of dilemma (contra Greene et al.'s [2004] claim). Instead, they represent differential demands within the same network that depend on the specific stimuli used. Given this view of the whole network, we can then drill back down into the function of specific areas to see how they look.

That kind of abduction will probably only go so far, however. A better, more general approach will surely involve data-driven meta-analyses, the goal of which is to see which task contrasts really result in discriminable brain networks. Early work on this sort of analysis is promising. Poldrack and colleagues (2009), for example, have taken an explicitly data-driven approach, using a variety of factor analysis to show how different patterns of brain activity load on hypothesized dimensions of cognition. The factor loadings revealed a number of surprising similarities and distinctions between task categories.

Others have taken up task validation more directly. Lenartowicz et al.'s recent meta-analysis examining the construct of 'cognitive control' is an excellent example (2010). Competing theories of cognitive control have hypothesized a number of different constructs that might be involved in control tasks: response selection, response inhibition, task switching, and working memory. Lenartowicz et al. first collected activation from all papers in the BrainMap Database that contained contrasts meant to isolate any of the above constructs. They constructed a probabilistic map for each voxel that showed how likely it was that a particular contrast activated a particular anatomical region. Using a classifier-based analysis, they then determined empirically the discriminability of the brain patterns associated with each pair of constructs.

The results were informative. Most of the constructs were associated with brain patterns that were relatively distinct. Further, all were easily distinguished from bilingual language tasks, included as a control condition.

https://doi.org/10.1086/667843 Published online by Cambridge University Press

However, task switching performed poorly: the patterns associated with it were significantly less discriminable from any of the other constructs (Lenartowicz et al. 2010, fig. 4). Lenartowicz et al. discuss several potential reasons for this failure of discriminability. However, the data at least suggest that task switching may not belong in a good cognitive ontology and that such data-driven meta-analyses provide a reason to delete it from our theoretical toolbox (2010, sec. 3). Note, importantly, that the patterns of activation associated with discriminable constructs also overlapped to a fair degree. This means that one cannot use the data to localize any of the hypothesized constructs to specific brain regions. While the overall patterns were discriminable, each construct itself probably depends on the joint activity of more basic components.

This sort of work is still in its infancy and depends on a number of methodological assumptions that need careful consideration. For reasons mentioned above, however, I suggest that it is a more promising style of analyzing NI data. To bring it all back around, it is true that brain regions are functionally specialized. But focus on the function of brain regions has been, I suggest, premature. First, we need to take a closer look at brain networks and figure out what tasks they are associated with. That has the potential to show that our existing categorizations of cognitive tasks is wrong and, if so, where we need to revise. Only once we have done this can we move with confidence back to individual brain regions and figure out what their distinctive contributions to cognition might be.

REFERENCES

Abutalebi, Jubin. 2008. "Neural Aspects of Second Language Representation and Language Control." *Acta Psychologica* 128 (3): 466–78.

Buckner, Randy L., and Daniel C. Carroll. 2007. "Self-Projection and the Brain." *Trends in Cognitive Sciences* 11 (2): 49–57.

Cohen, Laurent, Stanislas Dehaene, Lionel Naccache, Stéphane Lehéricy, Ghislaine Dehaene-Lambertz, Marie-Anne Hénaff, and François Michel. 2000. "The Visual Word Form Area: Spatial and Temporal Characterization of an Initial Stage of Reading in Normal Subjects and Posterior Split-Brain Patients." *Brain* 123 (2): 291.

Green, David W. 2003. "Neural Basis of Lexicon and Grammar in L2 Acquisition." In *The Lexicon-Syntax Interface in Second Language Acquisition*, ed. Roeland van Hout, Aafke Hulk, Folkert Kuiken, and Richard J. Towell, 197–218. Amsterdam: Benjamins.

Green, David W., Jenny Crinion, and Cathy J. Price. 2006. "Convergence, Degeneracy and Control." *Language Learning* 56 (Suppl.): 99–125.

Greene, Joshua D., Leigh E. Nystrom, Andrew D. Engell, John M. Darley, and Jonathan D. Cohen. 2004. "The Neural Bases of Cognitive Conflict and Control in Moral Judgment." *Neuron* 44 (2): 389–400.

Haxby, James V., M. Ida Gobbini, Maura L. Furey, Alumit Ishai, Jennifer L. Schouten, and Pietro Pietrini. 2001. "Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex." *Science* 293 (5539): 2425–30.

Ishai, Alumit, L. G. Ungerleider, A. Martin, J. L. Schouten, and J. V. Haxby. 1999. "Distributed Representation of Objects in the Human Ventral Visual Pathway." *Proceedings of the National Academy of Sciences* 96 (16): 9379–84.

https://doi.org/10.1086/667843 Published online by Cambridge University Press

Kanwisher, Nancy. 2000. "Domain Specificity in Face Perception." *Nature Neuroscience* 3 (8): 759–63.
Kanwisher, Nancy, Josh McDermott, and Marvin M. Chun. 1997. "The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception." *Journal of Neuroscience* 17 (11): 4302.
Klein, Colin. 2011. "The Dual Track Theory of Moral Decision-Making: A Critique of the Neuroimaging Evidence." *Neuroethics* 4:143–62.
Lenartowicz, Agatha, Donald J. Kalar, Eliza Congdon, and Russell A. Poldrack. 2010. "Towards an Ontology of Cognitive Control." *Topics in Cognitive Science* 2 (4): 678–92.
Pessoa, Luiz. 2008. "On the Relationship between Emotion and Cognition." *Nature Reviews Neuroscience* 9 (2): 148–58.
Poldrack, Russell A. 2006. "Can Cognitive Processes Be Inferred from Neuroimaging Data?" *Trends in Cognitive Sciences* 10 (2): 59–63.
Poldrack, Russell A., Yaroslav Halchenko, and Stephen J. Hanson. 2009. "Decoding the Large-Scale Structure of Brain Function by Classifying Mental States across Individuals." *Psychological Science* 20 (11): 1364–72.
Price, Cathy, and Karl Friston. 2005. "Functional Ontologies for Cognition: The Systematic Definition of Structure and Function." *Cognitive Neuropsychology* 22 (3): 262–75.
Ramsey, Joseph D., S. J. Hanson, C. Hanson, Y. O. Halchenko, R. A. Poldrack, and C. Glymour. 2009. "Six Problems for Causal Inference from fMRI." *Neuroimage* 49:1545–58.

https://doi.org/10.1086/667843 Published online by Cambridge University Press