Brain Regions as Difference-Makers

Colin Klein
Macquarie University
colin.klein@mq.edu.au

Abstract

Contrastive neuroimaging is often taken to provide evidence about the localization of cognitive functions. After canvassing some problems with this approach, I offer an alternative: that neuroimaging gives evidence about regions of the brain that bear *difference-making* relationships to psychological processes of interest. I distinguish between the specificity and what I call the systematicity of a difference-making relationship, and I show how at least some neuroimaging experiments can give evidence for systematic difference-making.

1 Localization and its Discontents

Say that a cognitive function is *localized* when a distinct brain region performs that function and only that function. The idea that brain regions localize cognitive functions is a widely held one. The functional differentiation of the cortex is almost universally accepted. Complex cognitive functions will require many brain regions working in concert (Petersen and Fiez 1993), but at *some* level of decomposition it seems reasonable to expect a one-to-one mapping between cognitive functions and brain regions. That is arguably the simplest picture of how mind and brain are related, and so a good starting point for cognitive neuroscience.

Cognitive neuroscience thus has a strong commitment to the methodological strategy of *localization*: that is, of finding where in the brain cognitive functions are localized. Suppose, by way of example, we care about the ability to recognize objects. Cognitive neuroscience starts by decomposing this ability into a variety of sub-personal capacities, which collectively work in concert to implement that personal-level skill. Recognizing objects might involve several such processes: low-level visual processing, feature extraction, and categorization. Localization suggests that each of these subprocesses be assigned, on the basis of experimental evidence, to a particular brain region (or if not, that some further decomposition can).

Object categorization, for example, appears to localize to a region of the inferior temporal cortex (IT) (Ishai et al. 1999; Kriegeskorte et al. 2008; Kriegeskorte, Mur, and Bandettini 2008). That is based in part on evidence about differential neural activation between object recognition tasks and simple visual tasks. Localization gives a framework within which to interpret differential activations. To localize object decoding to the IT is to say that the IT is the *region for* object recognition. That is what IT *does*; that is its function. In one stroke we thus learn something both about how object recognition is implemented in IT and what a particular region of the brain does.

In philosophy of mind, localization fits well with the strategy of *homuncular functionalism* (Dennett 1981; Lycan 1981; Drayson 2012). Proposed to avoid explanatory regress in cognitive science, homuncular functionalism advocates decomposing complex cognitive mechanisms into simpler and simpler operations, until simple operations can be uncontroversially identified with neural processes. Philosophers of science have further emphasized the role of localization in

picking apart complex mechanisms in many sciences (Bechtel 2002; Craver 2007; Bechtel and Richardson 2010).

Finally, localization is often invoked as the rationale behind functional neuroimaging (NI). Functional neuroimaging provides evidence about brain activity. Traditional cognitive theories are couched in computational terms, and so make no prediction about brain activity (Coltheart 2006). A natural way to align them is to assume that cognitive functions are localized (Roskies 2007). That assumption lets us interpret activity in response to cognitive tasks as a guide to neural localization, and vice-versa. Localization thus tells us why NI is worth doing, and gives a framework within which to interpret NI evidence. Indeed, so close is the association between neuroimaging and localization that, in the early history of the technique, attacks on localization were mostly limited to neo-behaviorist (Uttal 2001) or strongly connectionist (Hardcastle and Stewart 2002) authors---that is, to authors who are generally skeptical about the project of cognitive neuroscience itself.

Yet localization has come under increasing empirical pressure from the very neuroimaging results it was supposed to help us interpret. Recall that localization requires that each brain region performs one, and only one, cognitive function: at some level there must be a one-to-one mapping between cognitive operations and brain regions. Yet there is now substantial evidence that cognitive functions simply do not line up neatly in this way, at any level of decomposition (Price and Friston 2005; Poldrack 2006).

Consider a brain region like inferior temporal cortex (IT), which is consistently activated in studies of object recognition. Numerous studies have shown IT activation in essentially the same location for very different cognitive processes.[1] Villarreal et al. (2008) find activation in the same area for visual recognition of *gestures* as opposed to objects. IT is marked as part of the "extrastriate body area" by de C Hamilton et al. (2006), who cite its activation in action judgments. Complicating the picture further, Sabri et al. (2008) note that this area is activated specifically by attended (as opposed to unattended) speech, showing that what it does is not confined to *visual* tasks. And finally, Goel and Dolan (2001) note its activation in abstract

[1]I take as representative peaks from Ishai et al. (1999) and Clarke and Tyler (2014). Studies found using Neurosynth, http://neurosynth.org, described in (Yarkoni et al. 2011).

(versus concrete) reasoning tasks, which does not seem to have much to do with perceptual classification at all.

Similar results are found throughout the brain (Anderson 2014). Conversely, there appears to be quite a bit of variability in how tasks are performed, both between individuals and within individuals at different times (Friston and Price 2003; Figdor 2010). These together suggest that the mapping from cognition to brain is many-to-many.

It remains unclear whether localization will survive this challenge. Some authors, following Price and Friston, take the problem to lie in the "cognitive ontology" that underlies traditional cognitive science. They view failures of localization as evidence that we need to revise our traditional cognitive theories, to a more or less radical degree (Price and Friston 2005; Poldrack, Halchenko, and Hanson 2009; Anderson 2015). Others suggest that traditional cognitive categories might be maintained and localized so long as we give more sophisticated descriptions of regional functioning (Roskies 2007; Rathkopf 2013; Hutzler 2013). Still others suggest a move to context-sensitive localizations that take into account things like the task being performed or the wider network in which regional activation can be embedded (McIntosh 2004, Klein 2012, Burnston 2016). Each of these preserve something that looks like localization, but at the expense of significantly complicating the simple picture with which we began.

It is worth taking a step back. Localization is an interesting relationship when it can be found, and may play a role in cognitive neuroscience. But cognitive functions don't have to be localized, and cognitive neuroscience can proceed without the assumption of general localization. Functional neuroimaging can also be interpreted usefully under a weaker set of assumptions. That alternative is worth taking serious.

This paper sketches such an alternative way of thinking about brain regions: as things which *make a difference* to cognitive function. The presentation comes in three steps. First, I sketch out the difference-making relationship as it holds between brain and cognition. Difference-making is a weaker metaphysical relationship than localization, and compatible with the diversity and complexity of NI results. Second, I argue that contrastive neuroimaging can be understood as giving evidence about the difference-makers for psychological processes. Thus by treating brain regions as difference-makers, we get a unified framework within which to treat NI evidence.

That does not yet say, however, how *good* NI evidence is, particularly in the fraught context of adjudicating between cognitive and psychological-level theories. So, third, I will argue that at least some contrastive neuroimaging techniques (though not, perhaps, the historically most common ones) can give relatively high-quality evidence. This is an additional virtue of the difference-making account: it allows us to make distinctions *within* types of neuroimaging evidence. I conclude with a worked example.

2 Brain Regions as Difference-makers

Return to the role of IT in object recognition. Whether or not object recognition can be localized in IT, we might get evidence for something else: that by *manipulating* IT, one could manipulate object recognition.

Suppose you look at a picture of a camel, and recognize it as such. To say that IT is a difference-maker for object recognition is to say that if we could intervene on IT (in the sense outlined by Woodward (2003)), it would be possible to change some fact about how you recognize objects. That change might be to whether you recognized anything at all, to *what* you recognized, to how quickly you recognized it, or to your chances of making an error when you recognized it. Note that IT might only allow us to manipulate a handful or even only one of those parameters. That is enough to make it a difference-maker for object recognition.

More formally, to say that a region R makes a difference to a subject's f-ing (against a background of brain activity M) means that:

1. There is a set of ways of f-ing $\{a_1 \dots a_n\}$ that vary along some dimension (i.e. in the presence or the manner of f-ing)
2. There is a set of distinct ways $\{b_1 \dots b_n\}$ that R could be active
3. When a subject fs in manner a_n, R is active in one of the ways b_n, and
4. If one were to intervene on R such as to change b_n to some other b_i while keeping (at least initially) M the same, the subject would have done one of the other a_i instead of a_n.

Condition one requires some meaningful sense in which an activity f can vary. The simplest case is just the contrast between f happening or not. But recognizing a camel can also happen quickly or slowly, confidently or hesitantly, and so on. Further, what I recognize can be a camel *rather than* a car or a clown; that is, object recognition can vary in content. No dimension of variation is privileged, though some will be more useful. I will model these a_ns as determinates of the determinable F. Since f-ing will typically be variable along several dimensions, the same region might be able to make a difference in more than one way. When we consider interventions on f-ing, the ways of varying f must be well-defined. (As they are in all interventionist accounts; see Woodward (2003, 115ff).)[2]

Condition two ensures that there is a similarly well-defined sense in which regional activation can vary (including its absence). As this is an interventionist account, the important cases will be ones where R is made to vary while keeping the rest of brain activity fixed (at least initially). Such interventions need not be within the realm of current or even ideal human technology; the point is merely that *some* intervention is possible. Finally, the distinct ways of varying activation in R might form different sets for different ways that f might vary.

Conditions three and four capture the idea that changing the way in which R was activated would make a difference to what happened with the subject's f-ing. What it means to say that IT makes a difference to object categorization, then, is that when I recognized the camel there was a

[2] There is, properly speaking, no variation *between* tasks on such a picture, simply variation within tasks. In that regard, the present picture resembles Sternberg's (2011) account. However, my account is more permissive than Sternberg's. The determinable-determinate structure of tasks is (in many cases) hierarchical, and so a region could make a difference between task determinate which themselves determinables of a higher-level determinate. To take a fictional example, if recognizing objects and recognizing letters are ways of recognizing *things*, then we can meaningfully ask which brain regions make that difference. Conversely, when tasks are not plausibly related in that hierarchical way, contrasts between them don't make much sense in the first place: you could try to contrast (say) dancing and reading, but it's not clear why that would be a good idea.

pattern of activity in IT, and were scientists able to alter the activity just so, I would have (wrongly) identified what I saw as a car instead (or reacted more slowly, or less accurately, or…). I leave further details about the specifics of R open. Many different brain parameters might be difference-makers, and different neuroimaging techniques will deliver information about different kinds of neural parameters. It is an advantage of the account to be pitched abstractly enough to handle any of these. Finally, the set $\{b_1 \ldots b_n\}$ needn't cover all ways that R might be active: there might be further, f-irrelevant changes possible. The important thing is that there is *some* set of manipulations of R that could give us a handle on the subject's f-ing.

Treating brain regions as difference-makers in this way has a number of metaphysical advantages over localization. First, difference-making is a weaker metaphysical relationship than localization. Localization entails difference-making: manipulating the region where a function is localized is surely a way to manipulate parameters of that function. Difference-making does not entail localization. A brain region might make a difference to a process without being the place where that function is localized. The function may not be localized at all. Hence difference-making may still be true when localization is not.

Second, the difference-making account is more flexible than the localizationist account. Activity in IT makes a difference to at least three different things. First, it makes a difference to *seeing objects*: a personal-level process, performed by a whole agent. Activity in the IT also makes a difference to *object category decoding*: a subpersonal-level cognitive process, which combines with other subpersonal processes to explain seeing objects. And IT makes a difference to *other brain regions*: that is, to other processes described at the neural level and that bear some (here unspecified) implementation relationship to the subpersonal. All three of these difference-making relationships are interesting, and potentially theoretically important for cognitive neuroscience. By contrast, it is unlikely that personal-level processes can be localized to brain regions.[3]

[3]Though *contra* Bennett and Hacker (2003), I do not think that localizing personal-level processes is a *conceptual* confusion. The cruder phrenologists might have been correct, and personal capacities might have been localizable as well. We just have very strong empirical evidence that personal-level processes do not localize, and that localization itself will be of "elementary operations" (Petersen and Fiez 1993).

Third, difference-making is compatible with the empirical evidence that made localization problematic. The many-many mapping between cognition and the neural was a serious challenge to localization—it seemed that either we must abandon localization or contemplate considerable rearrangement of our current cognitive categories. Difference-making is more accommodating. On the face of it, there is nothing odd about the fact that IT makes a difference to many different psychological processes. Many different brain regions (as I will show shortly) also make a difference to object recognition.

A useful parallel move (which has influenced me) has been made about genetic function. As Sterelny and Kitcher put it, talking about genes for eye color:

> Nobody who subscribes to this practice of labelling believes that a pair of appropriately chosen stretches of DNA, cultured in splendid isolation, would produce a detached eye of the pertinent color. Rather, the intent is to indicate the effect that certain changes at a locus would make against the background of the rest of the genome. (Sterelny and Kitcher 1988, 349).

To say that "G is a gene for t," then, means that varying the details of a gene at a particular locus would lead to variation in t. This is compatible with a gene being a difference-maker for other traits, and many things making a difference to that trait. Difference-makers are usually assumed to operate against a causal background. They do not compete with each other for the title of *the* thing responsible for the outcome. So too with brain regions. Difference-makers are theoretically useful precisely because they do not compete with one another for causal influence.

3 Methodological Advantages of Difference-making

A difference-making account also provides an interpretive framework for the evidence provided by contrastive neuroimaging. All sorts of neuroimaging, insofar as they provide evidence about regional brain function, can be accommodated within the difference-making framework.

I will be particularly concerned with *contrastive* uses of neuroimaging. Contrastive neuroimaging includes any use of any imaging modality which takes as primary data the difference in brain activation between two or more conditions of interest. Traditional subtractive designs are contrastive. So are designs where one or more factors are varied parametrically. So are most multivariate (or 'decoding') analyses. Each relies on task contrasts, and differ in their analyses of the data arising from those contrasts.

The primary evidence in contrastive designs is a set of brain regions in which activity differed in some way between tasks of interest. Initial studies of IT were paradigmatic contrastive studies: recognizing objects is compared with (say) fixating on matched visual noise, and the result was a brain region (the IT) that was differentially active between those two conditions.

We may naturally interpret contrastive studies as showing brain regions that are difference-makers between the two tasks involved. The reasoning is straightforward. Physicalism demands that a difference in task performance must ultimately be traced to a difference in brain activity somewhere along the line. Recording differential activity is likely to pick out the regions that made that difference. This is not knock-down evidence, and some sorts of evidence are better than others (a point to which I will return). Nevertheless, contrastive activity seems like a defeasible guide to at least some of the actual difference-makers for the contrast in performance.[4]

Here, a potential confusion should be addressed. Many studies set up the task contrasts between *stimuli*. Obviously, brain regions do not make a difference to causally upstream events like stimulus presentation. In our running example, IT does not make a difference to whether the subject was *presented with* a camel or a car. Rather, IT makes a difference to whether a subject *perceived* a camel or a car – that is, to the cognitive process evoked by the stimuli.

This interpretation also brings certain pragmatic advantages. The results of contrastive studies are often presented in an absolute way—e.g., one might report that "IT is active when subjects recognize objects." This is misleading in a number of respects. IT is not quiescent during other tasks; the whole brain is constantly active, and task-related fluctuations are relatively small compared to this baseline activity (Raichle and Mintun 2006). This baseline is conceptually important, and obscuring it risks giving a misleading picture of brain function (Klein 2014). Difference-making accounts, by contrast, more accurately represent the *differential* nature of task-related responses.

Further, the contrastive nature of studies is important when there is *disagreement* about cognitive theories. For different theories might interpret different contrasts to show different things. Consider, for example, debates over the existence of a neural module for face processing. Kanwisher and colleagues took differential fusiform activity in face- versus house-viewing to

[4] For actual difference-making and its important scientific role, see (Waters 2007).

show the presence of a specialized module for face recognition (1997; 2000). Haxby and colleagues disagreed: they took that differential activity to show the parts of a general distributed network that happened to work harder during face recognition (2000).

This disagreement is in part about how to interpret the task contrast itself: Kanwisher took viewing faces versus houses to be a contrast between two distinct domains of visual objects, whereas Haxby took it to be variation within a single domain. More formally: Kanwisher took the contrast to be one that provided pure insertion of a module, while Haxby took it to be a case of (poorly-defined) parametric variation within a single domain (for more on the distinction, see Sternberg (2011)). Hiding the contrastive nature of the study—by simply saying, for example, that "FFA is active when subjects look at faces"—is to omit information crucial for situating the dispute.

Difference-making accounts of contrastive neuroimaging, by contrast, do not suffer from this problem. Difference-making is intrinsically contrastive: to make a difference is to make a difference *between* two states of affairs. Similarly, it is straightforward to move from contrastive studies to difference-making claims: the difference made can be indexed to the contrasts used. This does not *solve* the problem, of course, but it does not obscure it either.

4 Difference-making and explanatory projects

I have said what sort of evidence contrastive neuroimaging can give, but not yet how that evidence might be useful. It is worth distinguishing two distinct projects the cognitive neuroscientist might have. Difference-making evidence is relevant to both, but is relevant in different ways.

On the one hand, there is the project of figuring out the information processing performed by a particular brain region. This is an exercise in figuring out the computational architecture of the brain at the spatial scales addressed by NI, with an ultimate eye to linking that story to smaller-scale computational neuroscience. On the other hand, there is the project of figuring out the cognitive architecture of the mind: that is, uncovering the traditional cognitive science project of discovering the computational mechanism that best explains observed psychological capacities and processes.

These projects are often not carefully distinguished. Perhaps this is because the localizationist picture does not really distinguish the two stories: looking at the computational architecture of the brain just *is* looking at the localized versions of the functions which make up the mind. Yet the relationship between high-level computational states and the low-level computational states which implement them can be complex and messy. That is where the localizationist ran into trouble. The complexity is not something special about brains, but about the complexity of computational implementation in the real world. As Jonas and Kording (2016) recently and elegantly showed, even very simple computer chips such as the 6502 microprocessor support a complex mapping between computational states and chip states.

Interpreting NI in a difference-making way, on the other hand, reveals two distinct roles for NI evidence. First, we can ask, "Why is it that this particular brain region makes a difference, in these ways, to these different psychological processes?" That takes as evidence the full breadth of cognitive relationships to which the region makes a difference. That in turn might demand a story about the computational function performed by the underlying brain region itself, even if that could not be cashed out in terms that could appear in any cognitive theory (Klein 2012).

Second, we might ask, of any set of putative personal or subpersonal capacities, "Why is it that the difference-makers for these processes converge and diverge in the ways that they do?" That is, we might use difference-making information in a *taxonomic* way, to tease out the category structure of a cognitive ontology. Looking at convergence and divergence of causal factors in order to delineate entities has a long history both in science in general (Salmon 1984) and in psychology in particular (Campbell and Fiske 1959).

Both of these projects will still play an important part in a mechanistic explanation of psychological capacities. Some connections are straightforward. Sorting out the underlying computational properties of neural hardware gives a foundation from which mechanistic explanations can be built. Bottom-up facts about difference-making similarly form one half of Craver's constitutive relevance criterion for mechanistic parthood (Craver 2007). Assuming that brain regions *are* at the appropriate grain to be mechanistic parts (an empirical question), NI can be a source of evidence about what the parts are.

Yet patterns of convergence and divergence compliment mechanistic explanation in another, possibly more important way. Mechanistic explanation seeks the parts which give rise to some particular process. That leaves open the question of how we pick out the relevant process in the first place---and, in particular, how we group and distinguish processes of the same type. Suppose all particular instances of object recognition are, in fact, the same type of process. That suggests that they require the *same* sort of mechanistic explanation. If recognizing camels and recognizing cars are in fact determinables of the same determinate, then the same mechanism ought to explain both. Yet traditional accounts of mechanistic explanation do not (I submit) say much about how we are to categorize and group together the processes that ought to have similar explanations.

Similarly, as we've seen, the very same lower-level item might have a flexible function that is useful in a variety of different contexts. Mechanistic explanations themselves say nothing about whether lower-level parts in one explanation are identical to the parts in another.

Mechanistic explanations, in sum, leave unresolved certain ontological questions about the taxonomy of the processes and entities they postulate. In the case of the brain, a difference-making perspective can weigh in on these questions. For by the mechanist's own lights, the chief criteria that ought to govern the taxonomies of mechanistic ontology ought to involve the sorts of interventions that it is possible to make on those parts and the structure of the effects that those interventions have. Insofar as neuroimaging might play a role in explaining the mind, then, it might be via this alternative, indirect route: not (just) by giving direct evidence about cognitive mechanism, but rather by structuring and categorizing the mechanisms that exist.

5 Ways of Making a Difference

The difference-making account is permissive by design. A great many brain regions turn out to be difference-makers for object categorization. Consider the reticular formation (RF), a brainstem region that controls (among other things) sleep-wake cycles. RF is a difference-maker for object categorization. Change its activation in the right way and you go from recognizing objects to not (because you go from conscious to unconscious). Or take an early visual area like V1. It can make a difference like RF—remove activation and you'll be blind, which means you cannot recognize visually presented objects. But V1 also contains roughly retinotopic

representations of what you see; it is a way-station on the way to more complex visual processing. When subjects categorized objects, we could have altered their V1 activity in such a way that they would have seen something different, and then categorized that different thing instead.

On my account, then, both RF and V1 end up regions for object recognition. This is actually the right result. Being conscious *does* make a difference to whether you see objects, as do facts about what you see. Nevertheless, you might wonder how *useful* it is to point this out. One might worry that evidence had too cheaply is no evidence at all. The challenge of the remaining sections, then, will be to use the resources of the difference-making literature to sketch a more fine-grained picture of the *useful* difference-making evidence that something like NI might provide.

Here, I think the framework of difference-making permits a nice move. As Woodward notes, most philosophical work on causation has focused on what makes something a causal relationship at all. But there is another important project, that of "elucidating and understanding the basis for various distinctions that we (both ordinary folk and scientists) make *among* casual relationships" (2010, 287). Localization is strong enough to be interesting in its own right. Once we move to thinking about difference-makers, it is no longer enough just to note that some brain regions are difference-makers. It also becomes important to think about *how* they make a difference.

With that in mind, I will detail two aspects of difference-making that distinguish IT as a more interesting region than RF or V1. The first distinction among causes is *specificity*, in Woodward's 'influence' sense of the term. A more specific cause is one that has more fine-grained influence over the effect. In general, a specific relationship between C and E is one where, by manipulating C into many different states, we can reliably manipulate E into many different states. As Woodward puts it, "C will influence E to the extent that by varying the state of C and its time and place of occurrence, we can modulate the state of E in a fine-grained way" (Woodward 2010, 305).

Specificity is a graded notion. At one end, we have what Woodward calls 'switch-like' influence. C has switch-like influence over E when E can only be manipulated into one of two states. In Woodward's well-known radio analogy, the on-off switch of a radio gives switch-like influence

over the station which is currently playing. No matter how many effective positions there are of the switch, the switch can only flip the radio between two states (off, and whatever station it happens to be tuned to). At the other end is 'dial-like' influence. C has dial-like influence over E just in case C can be in many states and there is a 1-to-1 relation between states of C and states of E. In Woodward's radio analogy, the tuning knob has dial-like influence over the station: each stop along the dial corresponds (more or less) to a different station.

In general, dial-like influence is a more interesting and more useful scientific relationship than the less specific sorts. Many things tend to have switch-like influence over a phenomenon of interest, including most background conditions. That is true for the brain as well. Manipulations of RF, and crude manipulations of V1, are difference-makers for object categorization. But they are not difference-makers in any distinctive sense. At best, they tell us that we need to be conscious, or able to see, to recognize objects. We knew that already.

IT, by contrast, has a more fine-grained influence over object recognition: different states of IT appear to correspond to distinct object categories (Carlson et al. 2014). That means we could vary not just whether you recognize objects, but *what* you recognize by varying the state of IT. That seems like a more informative relationship.

Specificity is useful, but there are further distinctions we need to capture. Consider the role of an early visual area like V1. As V1 roughly replicates early visual information, each difference between a seen object should correspond to *some* difference in V1 activation. So it should be possible to make more subtle manipulations of V1 that result in specific views of objects. Hence, more subtle manipulations of V1 may give us specific influence over which object you categorize. To capture the intuitive difference between V1 and IT, we need something more.

Return to Woodward's radio. The tuning knob, recall, has a specific relationship to the frequency to which the radio is tuned: there are many different dial positions, many different frequencies, and a (roughly) bijective mapping between the two. The knob has a tighter relationship to frequency than specificity, however.

As an illuminating contrast, consider the relationship between the selector knob on my washing machine and the temperature at which it washes my clothes. There are a number of different positions, individuated by type of fabric. However, the temperatures at which clothes are washed

are distributed in a haphazard way around the dial. 'Cotton' washes clothes at $60°C$; next to that is 'Eco Cottons,' which washes at $45°C$; then 'Delicates' at $30°C$; then 'Heavily Soiled' at $55°C$; and so on.

The selector knob bears a specific relationship to temperature. Individuating temperature in $5°$ increments, each knob position corresponds to a different temperature in the range of the washing machine, and each possible temperature corresponds to a different dial position. Like the tuning dial, then, there is a bijective mapping between positions and outcomes within the range.

Yet the tuning dial bears a further, more interesting relationship to frequency: there is a *systematic* relationship between dial positions and frequencies. Each turn of the dial has the same, predictable influence on tuning frequency. Not so with the selector knob: turning the knob one click can have widely differing effects on temperature. That makes the tuning dial better suited for controlling the variable to which it makes a difference: its systematic relationship to what it controls gives us, in some obvious sense, a better handle on the world.

Specificity and systematicity are conceptually related. Specificity is a measure of the relationship between states of the cause and states of the effect. Systematicity is a measure of the relationship between *changes* in the cause and corresponding *changes* in the effect. That is, *C* and *E* are related systematically just in case information about how much *C* changes carries information about how much *E* will change. (Further discussion of formal details can be found in the appendix.)

Similar criteria have long been recognized as important for the analysis of causal structures. In Hill's famous paper laying out the conditions on good epidemiological inference, for example, one important indicator of a causal relationship is the presence of what he calls a 'biological gradient' or a dose-response curve. As he notes,

> …the fact that the death rate from cancer of the lung rises linearly with the number of cigarettes smoked daily, adds a very great deal to the simpler evidence that cigarette smokers have a higher death rate than non-smokers. (1965)

He notes further that while it may be difficult to "secure some satisfactory quantitative measures of the environment" that would reveal such a relationship, "we should invariably seek it."

Reflecting on both Hill and radio knobs, John Campbell introduces the notion of a "control variable" (Campbell 2007; Campbell 2010; Weiskopf 2015). Tuning knobs and cigarette-smoking are control variables for radio frequency and the likelihood of cancer, respectively. Why? Precisely because they have a "large, specific and systematic correlation" with what they make a difference to (Campbell 2007). It is precisely these systematic relationships that make such variables useful for control—and control under intervention is arguably the very heart of the interventionist approach to causation.

Campbell enumerates several useful features of control variables: there ought to be a total function between states of C and of E; our cause variable must be phrased at the right level of grain, there ought not be "gratuitous redundancy" between states of C and the same E (though the function does not have to be one-to-one); the function linking C-states and E-states ought to be computable in practice; there ought to be a dose-response relationship; and C ought to be locally manipulable in order to affect E (Campbell 2010, 23–25). These criteria seem to me to be good, and (with some qualifications) probably necessary conditions on systematicity.

Yet to re-emphasize, the further important feature of specificity is the idea that a *change* in *C* will make a predictable change in *E.* Working with the intuitive idea, however, I think one can see how contrastive neuroimaging might give evidence for systematic relationships. I conclude by discussing a simple worked example.

6 Discovering Systematicity

Throughout, I have referred to inferior temporal cortex as one plausible difference-maker for object recognition. There are multiple converging streams of evidence for such a view (Logothetis and Sheinberg 1996; DiCarlo, Zoccolan, and Rust 2012). I want to consider a case where I think neuroimaging has provided particularly clear evidence for IT's role in object recognition.

The activity of V1 neurons in response to a particular stimulus determines a point in an extremely high-dimensional space. Distinct object categories will form regions in this space, each separated by a manifold—a continuous curved surface that splits the high-dimensional space. Crucially, those manifolds—even in V1 space—will not intersect. Instead, they will be

"tangled," folded together in high-dimensional space such that points on one side of the boundary might have most of its close neighbors on the other side of the boundary.[5] Perceptual categorization can be modeled as a process of successively "untangling" these manifolds, such that the implicit information becomes easier for the brain to use in decision-making (DiCarlo and Cox 2007).

So-called *representational similarity analyses* have proven useful in revealing which parts of the brain carry information relevant to object-categorization (Kriegeskorte et al. 2008; Kriegeskorte, Mur, and Bandettini 2008). In a representational similarity analysis, one first collects univariate, comparative statistics about the patterns of brain activation in a region. Each pair of activations is then further compared to one another, creating a *representational dissimilarity matrix* (RDM) that shows how similar or dissimilar the activity evoked by individual stimuli might be. Finally, the RDM data can be further dimensionally reduced, using techniques like linear discrimination analysis to perform multidimensional scaling. The results of this multidimensional scaling (MDS) show how different stimuli cluster together in the representational space of a particular neural area. Importantly, such analyses, when performed on IT, often show linearly separable clusters along dimensions like animacy/inanimacy, suggesting that category-relevant information is present and separable in IT. In early visual cortex, by contrast, the stimuli usually cluster along low-level visual features.

Such results are not conclusive, however. For one, although early visual cortex matches best with models of low-level visual features, it still has a relatively good match with models of object category (Kriegeskorte, Mur, and Bandettini 2008). This is to be expected if, as DiCarlo and Cox suggest, even early visual cortex contains all of the information relevant to object categorization, albeit in a difficult-to-exploit way (2007, 335). For another, the fact that a brain region caries information about object category does not yet show that it *represents* that information in such a way that manipulating it would make a difference to action. These are general problems with

[5]See DiCarlo and Cox (2007, 334) for the general picture. Note that this is an idealization; non-intersecting tangledness is only possible if each point in V1 space always corresponds to a unique object category in the division scheme, which is unlikely given the role of time-dependent effects on categorization.

univariate analyses of fMRI activity and the more complex analyses that are descended from them.

To overcome these limitations, Carlson et al. added a novel twist (2014). After performing a representational similarity analysis and finding the line that best distinguished animate and inanimate objects, they had subjects perform a categorization task on the same set of stimuli. The dependent variable was reaction time, and they hypothesized that distance from the separating line in the multidimensional scaling graph would predict reaction time.

That is exactly what they found. The reaction times were modeled well by a sequential probability ratio test, a very general and successful model of neural decision-making (Bogacz 2007). When applied to the linear distance from the MDS separating boundary, reaction times were well-accounted for; the reaction times in early visual cortex, by contrast, were poorly modeled (Carlson et al. 2014, figure 2).

This is, to use my preferred terminology, good evidence that IT bears a *systematic* difference-making relationship to object categorization. The representational space modeled by MDS defines a distance function, as does (trivially) the set of reaction times, and these have the right sort of relationship to one another. Further, the use of the sequential probability ratio test suggests a (widely instantiated) neural mechanism that makes this systematic relationship non-arbitrary. By contrast, though V1 might bear a *specific* relationship to facts about object categorization, it does not bear a systematic one (or, more precisely, it does not bear *this* systematic one).

Of course, systematicity cannot always be attained: sometimes the world is just unsystematic. The relationship between (say) DNA and proteins that are transcribed from it is not a systematic one. It is possible for neural relationships to be unsystematic. Suppose knowledge of particular faces was encoded by individual 'grandmother cells,' one per face (Gross 2002). Then there would not be a systematic relationship between faces and their neural signatures, only an arbitrary (but very specific) mapping.

So systematicity is not a necessary condition on being a good causal relationship. Some disciplines might be able to do without it. That said, many natural relationships *are* systematic, and when they are the details of that systematicity are important for teasing out the underlying

causal structure. Neural systems are a particularly good candidate for systematicity, I submit. Grandmother cell-style coding is profligate; most modern theories of representation stress something like the systematic relationship between represented properties and a sparse, compact neural code.

Further, if systematicity *is* in place, it gives us a very useful set of principles by which to manipulate the variables we care about. That is crucial on difference-making accounts of explanation, in the same way that specificity is crucial. For again, in difference-making we care about the degree of control a variable gives us over a system, and systematic difference-making gives us much better control.

Systematicity also seems to be both present and theoretically important in the brain. There are a number of spatially organized maps in the brain: retinotopy in V1, somatotopy in primary sensory cortex, and so on. Given that, I think there might be a straightforward sense in which the IT has a more systematic relationship to object recognition than V1. V1, recall, is roughly retinotopic. The same object can appear in a variety of places in the visual field, in a variety of fonts and orientations, and so on. This makes the relationship between V1 and objects relatively unsystematic (even if it is specific): activation patterns representing different objects are close together in V1-space, while patterns representing the same object can be far apart. IT—or *some* region, at least—takes this complexity and reduces it down to a more orderly code, one that allows us to recognize the same object no matter where or how it appears.

Systematicity and specificity together thus give us an intuitive way to carve up difference-making in the brain. IT, V1, and the reticular formation are all regions for object recognition. But within regions for object recognition, we can distinguish between those that give us insensitive or disorderly manipulable relationships and those that give fine-grained, orderly ones. Object recognition is just an example, of course. The same sort of story can be applied to any activity, and to any brain region that makes a difference to that activity.

Finally, I gave an example of where contrastive neuroimaging really does seem like it gives us evidence about systematic difference-making in the brain. At the very least, this ought to help us interpret brain activity in light of our theory of object recognition, and so provide a guide to the sorts of computations that IT performs (no matter what else IT does).

Switching to a difference-making account gives us a sense of brain regions that fits with empirical data. It also gives us a vocabulary with which we can talk about more fine-grained differences and similarities between regions. And those fine-grained orderly differences might, at a first pass, give us a better clue to what is going on in both the brain and the mind.

Appendix

A formal treatment of systematicity brings out several interesting properties. Given a representation of C and E in metric spaces and distance functions Δ_c and Δ_e that measure the distance between distinct states of C and E (respectively) in those spaces when C is intervened upon, then the relationship between C and E is *systematic* to the degree that Δ_c carries information about Δ_e.

Specificity and systematicity dissociate. The main text gave a case of specificity without systematicity. Systematicity occurs without specificity when there is a loose relationship between values of C and E but a tight relationship between changes in those values.[6] If there are many different causes of lung cancer, and many different factors that determine whether or not a cause is effective in bringing about cancer, the relationship might be relatively loose and degenerate. Nevertheless, even in such a case, an intervention that increases smoking rates will increase cancer rates in a predictable way. Or suppose that our radio has a certain amount of slow drift as it tracks frequencies, so that day to day particular dial frequencies vary. This drift reduces the specificity of the dial over time. Nevertheless, if the bias is consistent, then the same *change* in dial position will continue to have the same degree of effect on the tuned station.

The simplest systematic relationship holds when interventions on C have a linearly additive effect on E. The tuning knob meets this criterion: changes in the quantity on which we intervene (dial angle) have a simple, linear effect on tuning frequency. Linearity of effect is, of course, a frequently sought relationship in many kinds of studies. Systematicity shows why this is not merely a fetish. A nonlinear relationship between C and E makes the effect of an intervention

[6] More precisely, it can occur without systematicity in the Waters sense of 'specific actual difference making' (2007). Unlike Woodward's influence version of specificity, this notion is sensitive to the actual probability distributions of different causes (Griffiths et al. 2015).

crucially dependent on the starting point. If frequency depended on (say) the square of the dial angle, then the same intervention would have a different effect depending on the starting point. Such a relationship would be difficult to learn and to use reliably. In natural systems, linear relationships encourage a kind of modularity. Modular systems require consistent interfaces that make for well-defined effects of interventions regardless of the internal state of the module (Calcott 2014). Linearity of response lets one manipulate the present state of a system in a definite way without knowing the details of that state.

In simple cases, the distance functions Δ_c and Δ_e are effectively redundant: one could just use correlations between values of instances of *C* and *E*. The advantage comes when we move to more complex cases. First, the units of either distance function need not be the same as the units of C or of E. To take a common example from sensory physiology, many perceived quantities are related to the *logarithm* of the underlying physical magnitude that is perceived. Using a log scale shows that the relationship between input and perception is a systematic one, despite the relationship between the two being nonlinear (Hilbert and Klein 2014).

Distance functions might also show threshold effects. The limit case of a thresholded function will be a binary function that simply partitions changes into two categories of interest. A series of thresholds can be assembled into a step function. Many causes have a natural similarity structure such that distinct but similar causes give rise to the same effect. When this is the case, small changes in C up to a point might have no effect at all (because they preserve similarity), and then result in quite dramatic jumps (because they pass the threshold). (The tuning knob will be best represented by a thresholded step function. Most knobs have a bit of play, and then transition to the next station with perhaps a bit of intermediate gray area.)

Step functions are important for two reasons. First, they let us discretize what are otherwise continuous possibilities for change in one variable. This lets a naturally discrete quantity have a systematic relationship to a naturally continuous one. Without such discretization, many distinct changes in the continuous variable will correspond to the same degree of change in the discrete one, obscuring the underlying relationship.

Second, thresholded functions let us group like causes together, and distinguish them from dissimilar things. Many imaging arguments rely on demonstrating precisely this property: that is,

they show that all of the patterns that correspond to a particular state are more similar to one another than they are to patterns corresponding to distinct states.

Systematicity so considered is slightly more permissive than Campbell's requirement on control variables. Campbell suggests that we might "… require that neighboring distinct values of the cause variable should be mapped to neighboring distinct values of the outcome variable" (Campbell 2010, 24). Systematicity, as I have considered it, can be looser than this: distinct values with the same effect may be clustered together. Nevertheless, systematicity as defined preserves the overall intuition that a systematic relationship ought to group same with same, and like near like.

I have placed relatively few restrictions on the Δ functions. Indeed, I can think of only two *formal* restrictions on them: they must be monotonic, and they must be identical across the original ranges of C and E. The latter requirement simply follows from the working definition of systematicity that I have been using. The whole point of systematic relationships is that learning about changes in C gives you information about how much E has changed, independently of the original values of E and C.

Finally, one might be concerned that philosophical cleverness could make systematic relationships out of intuitively unsystematic relationships—perhaps with sufficient cleverness on the distance functions, or by gerrymandering to group variables in some useful but ultimately arbitrary way. I suggest that this is no longer a formal problem. Rather, any distance function must ultimately receive motivation from the underlying mechanism that connects C and E. For one, as on any interventionist account, there must be some independent reason for thinking that the grouping of variables is a good one. For another, there must be some independent reason for thinking that the similarities between C and E states, as set out by the distance function, are actually exploited by the mechanism that connects C to E. Note that this evidence need not require actually detailing the neural mechanisms involved, or (crucially) supposing such mechanisms to be localized **[suppressed for review].**

Acknowledgements

Thanks to Rosa Cao, Max Coltheart, Peter Clutton, Adrian Currie, Paul Griffiths, Annelli Janssen David Kaplan, Brendan Ritchie, Kim Sterelny, Karola Stolz, Dan Weiskopf, and Jim

Woodward for helpful discussions, and to audiences in Taipei, Canberra, Sydney, and Chicago for helpful feedback on earlier drafts.

Funding

This work was supported by the Australian Research Council under Grant FT140100422.

References

Anderson, Michael L. 2014. *After Phrenology: Neural reuse and the interactive brain*. Cambridge: MIT Press.

———. 2015. "Mining the Brain for a New Taxonomy of the Mind." *Philosophy Compass* 10: 68–77.

Bechtel, W., and R. C. Richardson. 2010. *Discovering complexity: Decomposition and localization as strategies in scientific research*. Cambridge: MIT Press.

Bechtel, William. 2002. "Decomposing the Mind-Brain: A Long-Term Pursuit." *Brain and Mind* 3: 229–242.

Bennett, Max R., and Peter MS Hacker. 2003. *Philosophical foundations of neuroscience*. Oxford: Blackwell Publishing.

Bogacz, Rafal. 2007. "Optimal decision-making theories: Linking neurobiology with behaviour." *Trends in Cognitive Sciences* 11: 118–125.

Burnston, D. C. 2016. "A contextualist approach to functional localization in the brain." *Biology & Philosophy*, 31(4):527–550.

de C Hamilton, Antonia F., Daniel M. Wolpert, Uta Frith, and Scott T. Grafton. 2006. "Where does your own action influence your perception of another person's action in the brain?." *NeuroImage* 29: 524–535.

Calcott, B. 2014. "Engineering and evolvability." *Biology & Philosophy*, 29(3): 293–313.

Camerer, C., G. Loewenstein, and D. Prelec. 2005. "Neuroeconomics: How neuroscience can inform economics." *Journal of Economic Literature* 43: 9–64.

Campbell, John. 2007. "An interventionist approach to causation in psychology." In *Causal Learning: Psychology, Philosophy and Computation*, ed. Alison Gopnik and Laura Schulz, 58–66. Oxford: Oxford University Press.

———. 2010. "II—Control Variables and Mental Causation." In *Proceedings of the Aristotelian Society*, 110:15–30.

Campbell, D. and Fiske, D. 1959. "Convergent and discriminant validation by the multitrait-multimethods matrix." *Psychological Bulletin*, 56:81–105.

Carlson, Thomas A., J. Brendan Ritchie, Nikolaus Kriegeskorte, Samir Durvasula, and Junsheng Ma. 2014. “Reaction time for object categorization is predicted by representational distance.” *Journal of Cognitive Neuroscience* 26: 132–142.

Clarke, Alex, and Lorraine K. Tyler. 2014. “Object-Specific Semantic Coding in Human Perirhinal Cortex.” *The Journal of Neuroscience* 34: 4766–4775.

Coltheart, M. 2006. “What has functional neuroimaging told us about the mind (so far)?” *Cortex*, 42(3): 323–331.

Craver, C. F. 2007. *Explaining the brain*. Oxford University Press, USA.

Dennett, D. C. 1981. *Brainstorms: Philosophical essays on mind and psychology*. Cambridge: The MIT Press.

DiCarlo, James J., Davide Zoccolan, and Nicole C. Rust. 2012. “How does the brain solve visual object recognition?.” *Neuron* 73: 415–434.

DiCarlo, James J., and David D. Cox. 2007. “Untangling invariant object recognition.” *Trends in Cognitive Sciences* 11: 333–341.

Drayson, Zoe. 2012. “The uses and abuses of the personal/subpersonal distinction.” *Philosophical Perspectives* 26: 1–18.

Figdor, Carrie. 2010. “Neuroscience and the Multiple Realization of Cognitive Functions.” *Philosophy of Science* 77: 419–456. doi:10.1086/652964.

Friston, K. J., and C. J. Price. 2003. “Degeneracy and redundancy in cognitive anatomy.” *Trends in Cognitive Sciences* 7: 151–152.

Goel, Vinod, and Raymond J. Dolan. 2001. “Functional neuroanatomy of three-term relational reasoning.” *Neuropsychologia* 39: 901–909.

Griffiths, Paul E., Arnaud Pocheville, Brett Calcott, Karola Stotz, Hyunju Kim, and Rob Knight. 2015. “Measuring Causal Specificity.” *Philosophy of Science* 82: 529–555.

Gross, Charles G. 2002. “Genealogy of the ‘Grandmother Cell.’” *The Neuroscientist* 8: 512–518.

Hardcastle, Valerie Gray, and C. Matthew Stewart. 2002. “What Do Brain Data Really Show?.” *Philosophy of Science* 69: 72-82.

Haxby, J. V., E. A. Hoffman, and M. I. Gobbini. 2000. “The distributed human neural system for face perception.” *Trends in Cognitive Science* 4: 223–233.

Hilbert, D. and Klein, C. 2014. “No problem”. In *Consciousness Inside and Out: Phenomenology, Neuroscience, and the Nature of Experience*, ed Richard Brown. Dordrecht: Springer.

Hill, Austin Bradford. 1965. “The environment and disease: association or causation?.” *Proceedings of the Royal Society of Medicine* 58: 295–300.

Hutzler, Florian. 2013. “Reverse inference is not a fallacy per se: Cognitive processes can be inferred from functional imaging data.” *Neuroimage* 84: 1061–1069.

Ishai, A., L. G. Ungerleider, A. Martin, J. L. Schouten, and J. V. Haxby. 1999 Aug 3. “Distributed representation of objects in the human ventral visual pathway.” *Proceedings of the National Academy of Sciences* 96: 9379–9384.

Jonas, E. and Kording, K. 2016. “Could a neuroscientist understand a microprocessor?” bioRxiv preprint. DOI 10.1101/055624.

Kanwisher, N., J. McDermott, and M. M. Chun. 1997. “The fusiform face area: a module in human extrastriate cortex specialized for face perception.” *Journal of Neuroscience* 17: 4302.

Kanwisher, Nancy. 2000. “Domain specificity in face perception.” *Nature Neuroscience* 3: 759–673.

Klein, Colin. 2012. “Cognitive Ontology and Region- versus Network-Oriented Analyses.” *Philosophy of Science* 79: 952–960.

———. 2014. “The Brain at Rest: What it is Doing and Why That Matters.” *Philosophy of Science* 81: 974–985.

Kriegeskorte, Nikolaus, Marieke Mur, Douglas A. Ruff, Roozbeh Kiani, Jerzy Bodurka, Hossein Esteky, Keiji Tanaka, and Peter A. Bandettini. 2008. “Matching categorical object representations in inferior temporal cortex of man and monkey.” *Neuron* 60: 1126–1141.

Kriegeskorte, Nikolaus, Marieke Mur, and Peter Bandettini. 2008. “Representational similarity analysis–connecting the branches of systems neuroscience.” *Frontiers in systems neuroscience* 2: 1–28.

Logothetis, Nikos K., and David L. Sheinberg. 1996. “Visual object recognition.” *Annual review of neuroscience* 19: 577–621.

Lycan, W. G. 1981. “Form, function, and feel.” *The Journal of Philosophy* 78: 24–50.

McIntosh, A. 2004. “Contexts and catalysts.” *Neuroinformatics*, 2(2):175–181.

Petersen, Stephen E., and Julie A. Fiez. 1993. “The processing of single words studied with positron emission tomography.” *Annual review of neuroscience* 16: 509–530.

Poldrack, R. A., Y. O. Halchenko, and S. J. Hanson. 2009. “Decoding the large-scale structure of brain function by classifying mental states across individuals.” *Psychological Science* 20: 1364–1372.

Poldrack, Russell A. 2006 Feb. “Can cognitive processes be inferred from neuroimaging data?.” *Trends in Cognitive Sciences* 10: 59–63. doi:10.1016/j.tics.2005.12.004.

Price, Cathy J., and Karl J. Friston. 2005. “Functional ontologies for cognition: The systematic definition of structure and function.” *Cognitive Neuropsychology* 22: 262–275.

Raichle, Marcus E., and Mark A. Mintun. 2006. “Brain work and brain imaging.” *Annual Review of Neuroscience* 29: 449–476.

Rathkopf, Charles A. 2013. “Localization and Intrinsic Function.” *Philosophy of Science* 80: 1–21.

Roskies, Adina L. 2007. “Are Neuroimages Like Photographs of the Brain?.” *Philosophy of Science* 74: 860–872.

Sabri, Merav, Jeffrey R. Binder, Rutvik Desai, David A. Medler, Michael D. Leitl, and Einat Liebenthal. 2008. “Attentional and linguistic interactions in speech perception.” *Neuroimage* 39: 1444–1456.

Sterelny, Kim, and Philip Kitcher. 1988. “The return of the gene.” *The Journal of Philosophy* 85: 339–361.

Sternberg, Saul. 2011. “Modular processes in mind and brain.” *Cognitive neuropsychology* 28: 156–208.

Uttal, William R. 2001. *The New Phrenology*. Cambridge: MIT Press.

Villarreal, Mirta, Esteban A. Fridman, Alejandra Amengual, German Falasco, Eliana Roldan Gerscovich, Erlinda R. Ulloa, and Ramon C. Leiguarda. 2008. “The neural substrate of gesture recognition.” *Neuropsychologia* 46: 2371–2382.

Waters, C. Kenneth. 2007. “Causes that make a difference.” *The Journal of Philosophy* 104: 551–579.

Weiskopf, Dan. 2015. “The explanatory autonomy of cognitive models.” In *Integrating Psychology and Neuroscience: Prospects and Problems*, ed. David Kaplan. Oxford: Oxford University Press.

Woodward, James. 2003. *Making Things Happen*. New York: Oxford University Press.

———. 2010. “Causation in biology: Stability, specificity, and the choice of levels of explanation.” *Biology & Philosophy* 25: 287–318.

Yarkoni, Tal, Russell A. Poldrack, Thomas E. Nichols, David C. Van Essen, and Tor D. Wager. 2011. “Large-scale automated synthesis of human functional neuroimaging data.” *Nature methods* 8: 665–670.