Two Kinds of Reverse Inference in Cognitive Neuroscience

Guillermo Del Pinal* and Marco J. Nathan†

To appear in Leefman & Hildt (eds.)
The Human Sciences after the Decade of the Brain, Elsevier

Abstract

This essay examines the prospects and limits of 'reverse inferring' cognitive processes from neural data, a technique commonly used in cognitive neuroscience for discriminating between competing psychological hypotheses. Specifically, we distinguish between two main types of reverse inference. The first kind of inference moves from the *locations* of neural activation to the underlying cognitive processes. We illustrate this strategy by presenting a well-known example involving mirror neurons and theories of low-level mind-reading, and discuss some general methodological problems. Next we present the second type of reverse inference by discussing an example from recognition memory research. These inferences, based on *pattern-decoding techniques*, do not presuppose strong assumptions about the functions of particular neural locations. Consequently, while they have been largely ignored in methodological critiques, they overcome important objections plaguing traditional methods.

Keywords: Reverse inference; cognitive neuroscience; multivariate-pattern analysis; mirror neurons; mind-reading; simulation theory; theory-theory

*Zentrum für Allgemeine Sprachwissenschaft (ZAS), Schützenstr. 18m D-10117 Berlin, Germany. Email: ged2102@columbia.edu

†Department of Philosophy, University of Denver, 264 Sturm Hall, 2000 E. Asbury Avenue, Denver, CO, USA, 80208. Email: marco.nathan@du.edu

1 Introduction

Cognitive neuroscience is based on the plausible idea that our understanding of the mind has much to gain from investigations into the workings of the brain. Over the last few decades, this steadily-growing field has substantially advanced our studies of relatively modular systems, like vision and touch, and promises to seriously contribute to the resolution of longstanding disputes in various domains of higher-cognition. To achieve this ambitious aim, one of the most commonly used techniques is *reverse inference*, that is, the practice of inferring, in certain tasks, the engagement of cognitive processes from patterns or locations of neural activation. Since different psychological theories often make incompatible assumptions about the processes underlying a specific cognitive task, reverse inference can, in principle, be used to discriminate between competing hypotheses.

Scientists and philosophers often talk about reverse inference tout court. However, this article shows that it is crucial two distinguish between two different types of reverse inference. In the first kind, cognitive processes are inferred from the particular locations of neural activation observed in particular tasks. We examine these location-based inferences through a case study on the nature of mind-reading. Some prominent scientists have argued that mirror neurons provide decisive evidence for embodied theories of mind-reading. Their argument is based on a paradigmatic location-based reverse inference. In the first part of this essay, we show that this argument fails (§2) and, in doing so, we highlight some inherent problems with this kind of inference (§3).

Critiques of location-based inference are widespread. Indeed, prominent researchers have gone as far as suggesting that reverse inference should be removed from the toolkit of cognitive neuroscience. In the second part of this essay, we maintain that this more radical step should be resisted. Drawing on a recent case study of recognition memory, we argue that a second kind of inference, based on pattern-decoding techniques, overcomes the problems faced by location-based inferences. In particular, we show that pattern-based inference does not presuppose any problematic 'neo-phrenological' assumptions about functional localization in the brain. As a result, pattern-decoding techniques overcome some of the oldest and most resilient objections which have been raised against the methodology of cognitive neuroscience (§4). Although pattern-based inferences are quickly gaining popularity among cognitive neuroscientists, they are still largely ignored in most methodological discussions.

2 Location-Based Reverse Inference (LRI)

To introduce *location-based reverse inference* (LRI), consider some neuroscientific studies of 'mind-reading,' the capacity to identify and predict the mental states and behavior of others. The two main competing cognitive explanations of this capacity are the *theory theory* (TT) and the *simulation theory* (ST).

(TT) According to the theory-theory, mind-reading is based on a science-like

folk theory of mind, which includes law-like generalizations over symbolic representations of the following categories: (a) observable inputs and mental states; (b) mental states and other mental states; (c) mental states and observable outputs (Gopnik and Wellman 1992; Fodor 1992).

(ST) According to the simulation theory, mind-reading is based on the capacity to take other agents' perspective by simulating their mental states and actions as if they were one's own. This allows us to access directly the intentional states or actions that, in our own case, would cause and result from the simulated states, and attribute them to others (Gallese and Goldman 1998).

In emphasizing the symbolic and abstract nature of higher-cognition, TT follows the cognitivist tradition of Descartes and Kant and, more recently, Chomsky and Fodor. In contrast, ST's assumption that the vehicles of cognition have a sensory-motor format, locates it in the empiricist tradition of Locke and Hume.

It should be clear that TT and ST provide distinct, intuitively plausible, and yet mutually incompatible accounts of mind-reading. How does one choose one over the other? Some authors have argued that *mirror neurons*—brain cells that fire both when an organism acts and when the organism observes the same type of action performed by someone else—provide decisive evidence in favor of ST over TT (Gallese and Goldman 1998; Iacoboni 2009; Rizzolatti and Sinigaglia 2010).[1] The key data was obtained using experimental variations of two tasks (Rizzolatti et al. 2009; Kilner and Lemon 2013). The first task involves *executing* a basic-level motor act, such as grabbing a cup, or more complex tasks, such as grabbing a cup to drink from it; the second task involves *observing* another agent performing the same type of motor acts. To see whether these discoveries provide evidence in favor of ST or TT, let us begin by spelling out the predictions, at both the cognitive and neural levels, of each hypothesis.

According to TT, subjects possess concepts for simple motor actions, such as GRAB. These simple concepts, conceived as a-modal symbolic representations, can be recursively combined with other concepts to represent more complex motor acts, such as GRAB A CUP TO DRINK, and invoked in folk-psychological generalizations expressing regularities such as 'subjects who grab cups to drink tend to be thirsty.'[2] The crucial point is that, according to TT, the same

[1] There are various reasons why studies of motor-acts and mirror neurons are especially apt to clarify the structure of LRI. First, the relevant neural-level data is quite clear and has been extensively replicated. Second, such data was obtained by using single-cell recordings, which allows us to focus on reverse inference *per se*, without addressing tangential problems regarding more controversial data-gathering tools, such as fMRI.

[2] Although these concepts are a-modal, they interface with sensory and motor processes. On the sensory side, these concepts can be applied based on perceptual cues; on the motor side, they can be used to form intentions to act so that, when we form an intention to *grab that red cup*, one usually executes the corresponding motor act and, indeed, grabs the red cup as opposed to, say, biting it. These interface conditions are sometimes called 'legibility constraints' since, to be usable, a conceptual system (even an a-modal one) must be legible at its input-output interfaces, so that perceptual inputs can lead to the tokening of concepts, and action concepts can be translated into the appropriate motor commands.

concept, GRAB, is tokened both when an act is categorized as a 'grabbing act' and when it is intentionally executed, that is, both when we categorize (or imagine) the grabbing actions of others and when we form an intention to grab. TT's prediction that both *execution* and *observation* tasks involve tokens of GRAB (or, in more complex cases, tokens of GRAB CUP∧DRINK) has implications at the neural level. If TT is correct, then the neural implementation—and, consequently, the neural activation pattern that codes for tokens of GRAB—should be instantiated when grabbing actions are both *executed* and *observed*. Of course, TT does not presuppose that such neural patters are identical; all it requires is that both representations be tokens of the same type, that is, tokens of the concept GRAB.

Next, consider ST. Recall that, on this view, simulation is the default mind-reading process. When subjects perceive others as grabbing a cup, they simulate this basic-level motor action as if they were executing it themselves. This simulation allows a subject to retrodictively determine the intention that she would have when performing that same act herself, in similar conditions.[3] The key point is that the process tokened in *observation* is a subcomponent (or is structurally analogous to a subcomponent) of the corresponding process tokened in *execution*. Hence, at the cognitive level, ST predicts that a subset of the execution-process is also tokened in observation-processes. At the neural level, this entails that the implementation of the part of the simulation which is shared with the execution of the motor process, should be instantiated in both *execution* and *observation*.

At this point, it should be clear that deciding whether TT or ST provides the correct explanation of mind-reading, based on the experimental contrast between *execution* and *observation*, is not as straightforward as it is sometimes assumed, as both theories predict that there is a key cognitive component present in both experimental conditions. According to TT, this is the tokening of an action concept; ST takes this to be the tokening of subsets of motor processes. Yet, if one could distinguish between tokens of motor-action concepts and tokens of motor-action simulations at the neural level, this data could be used to adjudicate between TT and ST. To see whether this can be done, we now consider the neural data.

Recall that the neural data was gathered in experiments that compared two basic conditions: in *execution*, subjects perform a basic-level motor act; in *observation*, subjects observe full or partial evidence that others are executing the same type of basic-level motor act. The neural data collected in these studies was obtained using single-cell recordings of macaques (Rizzolatti et al. 2009; Kilner and Lemon 2013), yet fMRI studies suggest that the basic findings

[3]The details of the simulation process can be spelled out in various ways. On one option, the process is rather direct, in the sense that the observed motor action is directly matched by a corresponding representation, which initiates the understanding of that action 'from the inside' (Rizzolatti et al. 2001). In more complex models, subjects generate a candidate goal for the observed action and then simulate such action as if it were their own. If the action matches the goal, then that is the goal of the perceived action; in cases of mismatch, the process can be repeated (Gallese and Goldman 1998). For our purposes, we can remain neutral between these two models.

also apply to humans (Rizzolatti and Craighero 2004). Neural-level analyses revealed both an *activation pattern*, and a *location pattern*. A group of mirror neurons, call it '*MN*,' is selectively activated in both execution and observation, and *MN* is localized in the premotor cortex.

Two clarifications are now in order. First, to say that some neurons are 'selectively activated in some task' implies that these same neurons are not differentially activated in other relevant tasks. This has three implications. (a) Mirror neurons are not differentially activated by acts that look similar to (but are not) motor acts, e.g., motions of a hand that are not instances of a grabbing action, but look relevantly similar (Rizzolatti et al. 1996). (b) Mirror neurons are activated by instances of the same type of basic-level motor act even if the cues vary in modality. To illustrate, mirror neurons are activated not only when a subject sees the full action of grabbing, but also when she sees only part of the action, or even if she hears evidence of a grabbing-action (Umilta et al. 2001). (c) Finally, the mirror neurons activated in the *execution* and the *observation* of a basic-level motor act, such as *grabbing a cup*, are also activated when the act is embedded in a more complex one, such as *grabbing a cup to drink*. Interestingly, if you compare a set of *execution vs. observation* conditions involving a complex act (grabbing a cup to drink) with a set involving a different complex act (grabbing a cup to clean), some mirror neurons fire in both sets of conditions, since both sets involve grabbing acts, whereas others fire in only one of the two (drinking vs. cleaning) sets (Fogassi et al. 2005; Iacoboni et al. 2005). Second, while we identified two different components of the result—the pattern across tasks and the location of the mirror neurons involved—ST-theorists rarely separate these two aspects. However, keeping patterns and locations distinct will allow us to determine the relative inferential weight carried by the specific location of mirror neurons. Importantly, both kinds of inferences—from patterns of neural activation to cognitive processes and from locations of activation to cognitive processes—are common in neuroscience. Neither of these kinds of reverse inference, however, should be confused with the *pattern-decoding techniques* discussed in §3.

With all of this in mind, we can now examine the reverse inference that allegedly supports ST over TT. Recall that, in reverse inference, the engagement of a cognitive process (in a set of tasks) is inferred from neural data, usually consisting of specific patterns or locations of neural activation. The conditional probability that a particular cognitive process is engaged, given a set of tasks and neural data, depends on the probability of the neural activation in the task, given the hypothesized cognitive process. Hence, in order to determine whether *MN* provides any reason to select TT over ST, or vice versa, we need to determine the likelihood of the neural results, given TT and ST, respectively.

Consider, first, the mirror neuron *pattern*, that is, the result that a particular set of neurons selectively fires at the same rate in both *execution* and *observation* tasks. As noted, TT assumes that both conditions engage tokens of the same motor concept. Consequently, TT predicts some neural overlap in both conditions, namely, the pattern that codes for these tokens of the same concept-type (in our example, the uniform and selective firing rate of *MN*). ST

makes essentially the same prediction but, in this case, the uniformity is due to partially overlapping motor processes being engaged in both conditions, as opposed to tokenings of the same concept.[4] In short, TT and ST both predict an overlap in part of the neural pattern observed in both conditions (*execution* and *observation*) but neither makes specific predictions regarding the fine-grained structure of this pattern.

Next, consider *location*, the result that *MN* (the set of mirror neurons that selectively fire at the same rate in both *execution* and *observation*), are located in the premotor cortex. These considerations seem especially relevant because of a key difference between TT and ST. According to TT, what *execution* and *observation* share is the tokening of the same concept, which is taken to be an abstract representation that cannot be reduced to sensory or motor processes. In contrast, ST claims that what both tasks share is a partial overlap of the same type of motor process, in one case as part of a motor act and, in the other, as part of a simulation. At this level of informal description, the location of mirror neurons might seem to provide decisive evidence in favor of ST over TT for, as ST-theorists note, most neurons in the premotor cortex are known to be involved in motor acts.[5] This suggests that mirror neurons implement subsets of motor plans rather than a-modal motor action concepts.

Despite the intuitive appeal of these considerations, we argue that ST does not predict or explain the location of mirror neurons better than TT; strictly speaking, both theories are equally compatible with the neural results. The problem with the above reasoning has two sources. First, it misconstrues the predictions of TT regarding the neural encoding of motor-act concepts; second, it overestimates ST's prediction regarding the neural implementation of simulations. We now elaborate both points, in turn. Beginning with TT, as noted, this hypothesis entails that, when planning a motor act such as grabbing a cup, agents form an intention that tokens the concept GRAB CUP. This intention then interfaces with and instructs motor regions that carry out the computations required to execute the action. In the observation case, the corresponding act is understood as an instance of GRAB CUP together with a representation of a different agent. While, strictly speaking, TT does not predict that tokens of

[4]Strictly speaking, TT and ST only make this prediction on the assumption that tokens of the same cognitive process have uniform neural implementations. Although, in principle, this uniformity assumption could be challenged (for example, based on radical forms of multiple-realizability), it is a fundamental presupposition of cognitive neuroscience that we shall take for granted in our discussion.

[5]Indeed, ST theorists sometime insist that it is the location of mirror neurons in the premotor cortex that provides decisive evidence in favor of ST over TT (Rizzolatti and Sinigaglia 2010). For instance, given that mirror neurons fire across modalities, one might be tempted to conclude that these brain cells are a-modal. ST theorists, however, resist this move and propose instead that, since mirror neurons are located in the premotor cortex, they should be conceived as translating various modalities to the motor modality, as opposed to translating modalities into a non-modal abstract representation. A further consideration often used to defend ST appeals to deficit patterns caused by neurodegenerative diseases that affect the motor system. However, once we admit that motor concepts could be tokened in and interface with premotor areas, much of this evidence becomes irrelevant for the debate between TT and ST (cf. Hickok 2014).

motor act concepts are encoded in premotor areas, such a discovery is perfectly compatible with TT. Given that premotor areas are involved in motor actions, this location is a very plausible candidate for the interface between tokens of motor intentions (involving motor act concepts) and the operations involved in motor executions. In short, although, strictly speaking, TT does not predict the location of mirror neurons, it is certainly not undermined by it either.[6]

This conclusion might be viewed as good news for ST. However, to conclude that ST is supported by the neural data, it is not sufficient to just show that ST is compatible with the MN location. In addition, we must also show that ST predicts that data in a stronger sense than TT. Does ST's prediction fit the neural data more naturally or accurately than the analogous prediction of TT? Unfortunately, the answer seems negative. ST is surely *compatible* with motor simulations being implemented in the premotor cortex; but the theory does not entail this result any more than TT does. To illustrate, suppose that *execution* and *observation* showed instead that the MN pattern was localized in premotor areas in one task but in the prefrontal cortex in the other. Would this undermine ST? Clearly not. The neural location that implements a simulation process could be different from the location that implements the simulated process for various reasons. For instance, this could be a hardware solution to getting and keeping the simulation processes offline. Of course, there would have to be a way to determine that the neural pattern is the implementation of a simulation of the target process, but this information could be revealed by the details of the actual firing pattern, independently of the location. Note that this sort of reasoning is not unusual. This is the way in which experimenters often try to show that, although the perirhinal cortex is not the locus of spatial processing, it carries spatial information, and although the hippocampus is not the locus of item-processing, it carries item related-information (more on this below).

Admittedly, questions concerning which areas of the brain *could* implement motor-action concepts and simulation processes are somewhat puzzling. The reason for this is that we do not know enough about how cognitive representations and operations are extracted, encoded, and tokened at the neural level to be able to substantially narrow the range of neural locations that could perform the categorization operations of TT or the simulation processes of ST. The conclusion of the above discussion is that TT and ST are both compatible

[6] This has interesting implications for the debate between 'a-modal' and 'embodied' theories of concepts. Nothing currently known about the nature of neural computation and representation prevents one from holding that a-modal concepts about, say, tactile, visual, or auditory domains are encoded in neural locations which are topologically close to the areas that process tactile, visual, or auditory stimuli. Thus, TT can assume, as a reasonable working hypothesis, that concepts for basic-level motor acts are encoded in areas topologically close to the motor areas involved in action execution. To be sure, this closeness between locations of concept tokenings and their corresponding input-output interfaces is not quite predicted by TT. This is because one cannot a priori dismiss other 'hardware' solutions to processes such as extracting conceptual categories from sensory and motor modalities, applying concepts to sensory inputs, and using concepts to form motor intentions that can interface with motor processes to execute actions. Still, TT does not presuppose any distance between sites where symbolic concepts are encoded (in long-term and working memory tasks) and the corresponding input-output processing regions with which they can interface.

with location-data provided by mirror neuron studies. To be clear, we are not denying that we know quite a bit about the function of premotor neurons, and by extension, of premotor mirror neurons, for instance, that they are 'involved' in action planning, execution and recognition. However, in debates between TT and ST what we have on the table are two competing accounts of the precise computational form of the processes involved in action planning, execution and recognition. Pointing out the premotor location in which those operations are implemented does not, by itself, help us select which of the competing operations are actually used.

3 LRI: General Problems and Limitations

In the previous section, we have seen that *MN location*, that is, the discovery that the relevant pattern of mirror neurons is localized in the premotor cortex, does not provide conclusive evidence to select ST over TT (or vice versa), as an explanation of mind-reading. In this section, we extend the discussion into some general limitations of location-based reverse inference (LRI). Specifically, we begin by identifying a key assumption for the proper use of reverse inference, which we call the 'linking condition'. Next, we argue that this linking condition is extremely hard to fulfil for any study that infers cognitive process from the location of neural activation. In the second part of the article, we present an influential technique, *multivariate pattern analysis*, supporting a different type of reverse inference, which overcomes the problems faced by LRI.

To spell out the assumption that, we maintain, any properly conducted reverse inference should meet, suppose that C and D are two competing hypotheses of the cognitive processes underlying some task t. Further assume that C posits the engagement of cognitive process c, D posits the engagement of cognitive process d, $c \neq d$ (c and d are not the same process), and let n stand for a differential pattern of neural activation in some specific location. A reverse inference from the presence of n in t to the engagement of c in t requires the existence of independent studies that establish a link between n and c. The reason for this should be obvious: in order for the observation of n to support C over D, there must be a corroborated connection between n and c, that is, it must be shown that n is evidence for c and that n is not evidence for d. Let us call these background studies 'linking studies.' Provided that there is a robust linking study connecting n with c, experimenters can use the observation that n is engaged in t to infer that c (and not d) is engaged in t, providing evidence in favor of C over D.

Important as they are, linking studies are hard to obtain. In particular, there are (at least) two problems that must be avoided. First, the connection between c or d with n cannot be obtained a priori, but must be discovered though painstaking experimental work. This experimental work typically requires determining whether n is connected to c or d, in the context of some task t^* which, to avoid circularity, must be different from task t.[7] However, if C

[7]To minimize the possibility of violating the linking condition, properly conducted reverse

and D differ not only in their predictions for task t, but also in their cognitive-level interpretation of t^* then the linking studies that associate n to, say, c are problematic, as they ignore one of the competing alternative theories, namely, D. Second, even if the linking studies of c and n are properly conducted, in the sense that they are not biased toward any alternative, the region of the brain where n is located could be multi-functional, that is, it might also be known to implement other cognitive processes. Obviously, the two problems are not independent: the less selective the brain region of interest, the stronger the chance that linking studies ignored that n could also implement d.

Let us now apply this reasoning to the debate between TT and ST. The appeal to *MN location* in support of ST over TT falls prey to precisely these shortcomings. Specifically, to give a functionalist-level interpretation of the *MN location*, ST-theorists do not appeal to any tasks in which it is reasonably uncontroversial that something like simulation processes (or some key subcomponent) is engaged. Furthermore, in the tasks that are considered—the variations of *execution* that generate mirror neuron activation—what is under dispute is precisely their fine-grained functional interpretation. This is a clear instance of the first problem isolated above: mere activation in premotor areas does not have a cognitive-level interpretation that is relevant to adjudicate between ST and TT. This is because ST and TT provide different cognitive-level accounts of the interface between intentions for and execution of basic-level motor acts. Consequently, they provide different accounts of what is going on in premotor areas, in tasks such as *execution*. As a result, obtaining the same activation in premotor areas in *observation* is compatible with both cognitive-level accounts: TT and ST.

It might seem reasonable to suppose that the mind-reading debate is particularly susceptible to these problems. Perhaps the differences between TT and ST, in the *observation vs. execution* experimental context, are especially subtle, or maybe the fine-grained computational diversity of the premotor cortex is unique. These conjectures, however, do not withstand serious scrutiny. More or less overt violations of the 'linking condition' can be found in a number of reverse inferences used in studies of higher-cognition (Coltheart 2006a, 2013). This is not accidental. The most common technique employed in studies of higher cognition to make neural data bear on competing cognitive-level hypotheses is LRI, where the engagement of a cognitive process in a task is inferred from a particular location of neural activation (Shallice and Cooper 2011). LRI faces a systematic difficulty in satisfying the linking condition—that is, avoiding the two problems identified above—because the brain regions of interest are seldom selective. Most neural areas implement several cognitive processes, ranging

inferences invoke, as part of their background studies, tasks that are relevantly different from those subsequently used to discriminate among the competing cognitive hypotheses. Furthermore, the links to neural data should be established in tasks in which experimenters can control, with reasonable confidence and without ignoring any of the theories that will be subsequently tested, the engagement of the relata on the cognitive side. Of course, in the tasks then used to evaluate the competing hypotheses, the engagement of the cognitive process of interest is at issue, and the probability of its presence is reverse inferred from the resulting location of neural activation.

from completely distinct and independent to subtly different and interconnected ones. It is this functional diversity, we surmise, that undermines the plausibility and robustness of many LRIs.

At this point, one could argue that the general lack of selectivity of brain regions does not undermine the plausibility of LRI. Rather it is a methodological 'caveat' that might also serve the positive function of a general warning against explaining serious scientific claims in terms of 'just so stories.' Indeed, recent discussions of reverse inference have promptly noted and addressed the problem that activation in a region could also signal the engagement of cognitive processes other than the ones posited by the hypothesis under scrutiny. Whether expressed in Bayesian (Del Pinal and Nathan 2013; Hutzler 2014) or likelihoodist terms (Machery 2014), if the experimental settings are designed appropriately and the linking studies are reliable, the general solution consists in recognizing that experimenters can ignore cognitive hypotheses which are not part of at least one of the competing theories. In short, while LRI is, in principle, a sound inference, its usefulness is much more limited than enthusiastic supporters often recognize. This is a problem since, illustrated in the mind-reading case, substantial debates often involve disputes about the fine-grained cognitive functions carried out in particular tasks and neural locations, and we need a way to discriminate between such hypotheses.

The tension between lack of selectivity and LRI occurs at various levels of analysis. At a relatively coarse level, the difficulty arises from the controversial assumption that brain regions are relatively selective for coarsely-defined processes. At a finer level, the difficulty arises from the computational diversity associated with single neurons or groups of neurons in a given region, as revealed by the discussion of mind-reading. Consider the fine-grained computational diversity of premotor areas such as macaque area F5. Some neurons in F5 are selective for action perspective, manner of approach, or final execution strategy. Some neurons are selective for type of goal. Some are selective for particular modalities and others are cross modal. In particular, some neurons are active *only during observation* and others only during execution (Kilner and Lemon 2013). Take a neural network with basic units with that amount of computational diversity, and consider how many different processes—some more simulation-like, others more categorization-like, and all mimicking the MN pattern—you could build out of those basic units. In particular, note that some neurons in premotor areas fire only during observation. Hence, it remains an open question how exactly we should conceive of the representational format of the ones that are mirror neurons, i.e., that satisfy the MN pattern. TT and ST provide two different hypotheses, equally compatible with the data, and both implementable in a location with this sort of fine-grained computational diversity.

In conclusion, although LRI is a sound inference, its correct application is too limited to be of general use in cognitive neuroscience. In the following section, we shall examine a different kind of reverse inference that, we argue, provides a more promising and widely-applicable technique to discriminate between competing cognitive-level hypotheses.

4 Pattern-Based Reverse Inference (PRI)

Consider, once again, the case of mind-reading. In order to properly assess the relative plausibility of TT and ST, one needs to set up experimental tasks in which the cognitive processes posited by ST (motor act simulation) are uncontroversially engaged, tasks in which the cognitive processes posited by TT (categorization) are uncontroversially engaged and, in each case, map the neural results onto the corresponding cognitive-level process. In a word, one needs the appropriate linking studies. One must then compare those neural-level results with the results obtained in *observation*, the task in which the identity of the underlying cognitive process is under dispute. By doing so, one can determine whether understanding the basic-level motor acts of others is more like a simulation process or more like a straightforward categorization process. In the previous section, we argued that *location-based reverse inference* (LRI) has substantial problems fitting this bill. The root of the trouble is the computational diversity of brain regions, which implies that evidence provided by rough activation in a specific area is of limited value. In this section, we present and discuss a different kind of reverse inference that, we maintain, fares better on this score.

Mappings between cognitive functions and patterns of activity in particular brain regions, even those based on single-cell recordings, are not sufficiently selective. This point was illustrated *via* canonical studies of mirror neurons, which provide a neural pattern, the *MN location pattern*, consisting of a set of neurons that fire at the same rate in *observation* and *execution*. As noted, this firing pattern has not been decoded, in the sense that we do not yet know what sort of process or content is encoded in such firing sequence: that simple pattern could implement a token concept for a motor act, a subset of motor processes, or various other states. However, there are other kinds of mappings that are much more selective. For instance, vector patterns of neural activity, which contain detailed information about fine-grained cognitive-level representations and processes, can be decoded with techniques such as *multivariate pattern analysis* (MVPA). MVPA uses tools from machine learning to create statistical machines—called 'classifiers'—which can decode the cognitive states or processes encoded in particular neural data sets, such as multi-voxel patterns obtained using fMRI. These decoded patterns can then be used to reverse infer the engagement of specific cognitive processes (Poldrack 2008, 2011; Poldrack et al. 2011; Tong and Pratte 2012).

To illustrate MVPA, consider a study in episodic memory research about the cognitive processes underlying our capacity to classify items as 'old' or 'new.' We formulate this recognition capacity as follows, where s ranges over 'normal' adults. A set E contains some items that are new to s and others that s has previously encountered. If s is randomly presented from an item $e \in E$ and has to determine whether or not she has already encountered e, s can reliably distinguish between 'old' and 'new' items. Most researchers now accept some version of a dual-process theory of recognition. Two prominent competing explanations are the following (note the interesting parallels with

the mind-reading debate):

(RR) Recognition decisions are based on two processes which draw on two distinct sources of information: *recollection* of specific details and non-specific feelings of *familiarity*. Recollection is used by default but, when such information is unavailable, subjects employ familiarity.

(RF) Recognition decisions are based on two processes which draw on two distinct sources of information: *recollection* of specific details and non-specific feelings of *familiarity*. However, neither is the default process: the source of information employed depends on *specific contextual cues*.

RR and RF posit the same components to explain recognition decisions; what distinguishes them are the different interactions among those constituents. According to RR, recollection information is used by default to determine whether or not an item is 'old,' and familiarity is only reverted to when such information is unavailable. In contrast, RF implies that certain contextual cues will sometimes induce subjects to make familiarity-based recognition decisions even when recollection-information is available.

To test these hypotheses, 'pattern classifiers' are trained to determine the specific multi-voxel patterns associated with recollection and familiarity processes. More precisely, classifiers are trained in tasks where experimenters are able to control which cognitive process is engaged, thereby explicitly meeting one of the linking conditions for reverse inference. For instance, in one experiment, which will serve as our main example, subjects were exposed to singular and plural words, such as 'shoe' and 'shoes' (Norman et al. 2009). These subjects were then scanned while performing recognition tasks involving previously examined items (e.g., a shoe) and unrelated lures (e.g., a bicycle). The recognition tasks were divided into two disjoint sets: *recollection blocks* and *familiarity blocks*. In recollection blocks, subjects were explicitly instructed to recall specific details of the mental image formed during the study phase, and to only answer 'yes' if they were successful in that recollection. In contrast, in familiarity blocks subjects were instructed to only answer 'yes' if they found the word familiar and to ignore any details they might recollect from the study phase. After a training phase, classifiers were able to reliably determine whether some multi-voxel pattern of neural activation is an instance of recollection or familiarity.

What gives MVPA-based inferences a substantial advantage over traditional LRIs is that the reliability of the classifiers can be established within the experiment itself. In our example, this can be done by saving a subset of the recollection and familiarity blocks for later testing (so that they are not used at the 'training stage') and then determining the rate at which the classifier correctly categorizes the corresponding neural patterns. This phase of the study, where experimenters can control which process is engaged, provides the links between recollection, familiarity and their corresponding multi-voxel patterns that can then be used in reverse inference. Once these links are established, one can test competing hypotheses RR and RF in cases where the engagement of

the sub-processes cannot be directly controlled.[8]

In a second phase of the study, subjects were scanned while being exposed to a mixture of previously observed items ('shoe' and 'ball'), unrelated lures ('horse' and 'box') and previously unobserved switch-plurality lures ('balls'). The subjects' task was to determine whether the words they encountered were 'old' or 'new.' To test the competing cognitive-level hypotheses, experimenters examined the subset of tested items for which subjects made correct recognition decisions. Note that, since these are cases where both recollection and familiarity information was available to subjects, *RR* and *RF* make different predictions about what is going to happen. According to *RR*, the classifier should categorize all the corresponding voxel patterns as recollections patterns, since this is the default. *RF*, in contrast, predicts a more variable classification, involving at least some instances of familiarity-patterns, given that neither pattern should be used by default. The MVPA experimental results support *RF* over *RR* (Norman et al. 2009). When both types of information are available, various contextual cues determine whether familiarity or recollection is used as the basis of a subject's recognition decision. In other words, contextual cues determine whether, according to the classifier, the multi-voxel patterns underlying recognition decisions resemble more unambiguous familiarity or unambiguous recollection patterns.

Let us call a reverse inference based on pattern-decoding techniques such as MVPA, a *pattern-based reverse inference* (PRI). We conclude our discussion by emphasizing three substantial advantages that PRIs have over LRIs, regarding both experimental practice and philosophical implications. First, PRI allows for the reliability of classifiers to be determined within a phase of the same experiment in which they are employed. This feature of PRI fulfills the linking condition and allows experimenters to quantify their confidence in particular reverse inferences, thus providing a substantial advantage over LRI. As discussed in §3, successful reverse inference presupposes the existence and availability of robust and accurate links between levels. The linking studies required to establish these bridge laws confronted LRI with several difficulties stemming, for the most part, from the lack of selectivity of most brain regions of interest. Consequently, when the differences between competing cognitive hypotheses are subtle, the linking studies often ignore or underestimate the resources available

[8]The question arises whether we can extend the reliability of classifiers obtained from the testing phase to cases in which the experiments cannot determine the engagement of the psychological variables, since the latter inevitably involve some variation on the task. There are various studies which suggest that classifiers perform well under task variations. For example, in one study pattern classifiers were used to predict phonemes. The classifiers were still successful when presented with data from voices which were not presented in the learning phase (Formisano et al. 2008). Hence, at least this much variation in the task does not affect performance. In a study of visual working memory, classifiers were trained on data elicited by unattended gratings, and then tested on whether they could also predict which of two orientations was maintained on working memory when subjects were viewing a blank screen. Again, their reliability was maintained despite the substantial difference in stimuli and tasks (Harrison and Tong 2009). Indeed, testing for this kind of robustness relative to stimuli/task variation is usually taken as evidence that the brain region from which the data was obtained really does provide information about the function of interest (Tong and Pratte 2012).

to one (or both) of these hypotheses. This problem becomes evident in the TT vs. ST debate, where the functional interpretation of premotor areas (especially in *execution* tasks) according to TT is basically ignored. To make things worse, even if the linking studies presupposed in a specific LRI are, in themselves, non-problematic, we still have to consider other tasks that might activate the brain regions of interest and ask, for each task, whether it might be relevant for the hypothesis under scrutiny. Readers familiar with meta-analyses of function-structure mappings know how frustrating and confusing these studies can be (Poldrack 2011; Lindquist et al. 2012). It is important to realize that the problem is not primarily due to the resolution of the neuroscientific tools commonly employed in LRI, such as fMRI. In this respect, the example of mind-reading, discussed in §2, is particularly instructive. The single-cell recording technique used in mirror neuron studies of macaques reveals quite starkly the fine-grained computational diversity of neurons in the premotor cortex. Hence, the main problem has to do with computational diversity, which goes all the way down to the function of single neurons, making it often difficult—or even impossible—to determine the precise degree to which an LRI should affect our confidence in a given hypothesis. In contrast, MVPA and other techniques employed in PRI, decode cognitive processes from multidimensional vector patterns—as opposed to regions—of neural activation. As noted, when using these techniques, the reliability of the classifier underlying a particular reverse inference can be established within the experiment. To be sure, different kinds classifiers will vary in their success depending, among other things, on the type of task, the number and specificity of learning trials, the brain regions used for analysis (if restricted), and the nature of the machine learning algorithms (Poldrack et al. 2011). In the above recognition example, classifier accuracy was around 60%. However, in other tasks, such as those involving categorization of basic-level objects, the accuracy of classifiers can be much higher (Haxby et al. 2014).

A second advantage of PRIs follows from the fact that classifiers do not 'assume' the functional localization of cognitive states or processes of interest. The multi-voxel patterns used by classifiers can be distributed across the brain. Of course, experimenters can restrict the analysis to particular brain regions, especially if there are independent reasons to believe that a specific brain area is involved in some task, or if what is under scrutiny is the degree to which a region is responsible for executing some task. However, classifiers can also take non-localized, widely-distributed multi-voxel patterns. This is especially useful when comparing complex multi-step processes that likely involve several brain regions. All the examples considered in this essay are of this complex kind, as are most models of higher cognitive capacities in general. The philosophical significance of this observation is that PRI does not fall prey to one of the oldest and most resilient objections against the traditional approach of cognitive neuroscience (including LRI), namely, that it assumes a strong and objectionable form of *functional locationism* (Nathan and Del Pinal 2015). Speaking directly against this misconception, studies applying PRI can shed light on the degree of modularity, specialization, and functional localization of various cognitive processes and brain areas. For instance, MVPA studies consistently show that

the ventral temporal cortex carries sufficient information for classifiers to accurately distinguish between animate and inanimate objects (Kriegeskorte 2011). A similar approach may also help in studies previously explored via LRI. To illustrate, in our discussion of mind-reading, we have seen that the pattern and location of mirror neurons provide too little information to discriminate between competing cognitive hypotheses TT and ST. As some authors have pointed out, this might be because the neural locus of cross-modal action understanding is more widely distributed, and likely includes non-motor areas (Spaulding 2012). This hypothesis can be explored with MVPA. Indeed, MVPA studies in this area show that most of the voxels used to reliably classify actions of the same type across modalities are distributed in areas that are *not* canonical human motor areas or homologues of the macaque F5 (Oosterhof et al. 2013).

A third, and final, advantage of PRI is that the information decoded from multivariate vector patterns does not depend on previous assumptions about the coarse function of a given brain region of interest. As discussed above, when using an LRI, assuming that some brain region has a specific function, coarsely identified, often does not help adjudicating between subtly distinct hypotheses. The TT vs. ST case provided a clear example, where the crucial difference lies between a motor simulation process and an abstract representation of motor goals or execution processes. The debate between *RR* and *RF* illustrated an even subtler case, where the key distinction turns on the dynamics (as opposed to the individual components) of recognition decision processes. Once again, single-cell recordings, such as those recording the firing of mirror neurons in macaques, are extremely instructive for they highlight the fine-grained computational diversity of neurons and neural networks within a given brain region such as the premotor cortex. If one were to build neural networks composed of units with that much computational diversity, one would be able to develop processes that resemble the simulations of ST or the categorization processes of TT. Consequently, even fine-grained activation of these areas in tasks of *execution* and *observation* cannot, by itself, be used to determine which of these processes is actually implemented at the neural level. As shown by the recognition memory example, PRI is especially well-suited to solve these kinds of problems. By training a classifier to 'learn' which multi-voxel patterns are associated, say, with paradigmatic simulation and abstract categorization processes, a PRI can detect whether these diverse computational micro-units are working together in a way that is more like simulation or more like categorization, in the context of a specific task (e.g., *observation*).

5 Concluding Remarks

Cognitive neuroscience has recently faced a theoretical, scientific, and popular backlash. Many philosophers and scientists have been consistently critical of the impact of neuroscience on the study of higher cognition (Fodor 1974, 1997, 1999; Coltheart 2006a,b, 2013; Tressoldi et al. 2012; Miller 2008). Similar concerns have also been raised in various popular publications (Satel and Lilienfeld 2013).

This reaction is not completely unjustified. For every article providing neuroimaging evidence against deontological ethics or classical cognitivism, there is another one arguing that neuroscience has taught us nothing of relevance about higher capacities such as decision making and language processing. Our discussion of reverse inference supports a measured optimism. Careful analysis of the role of mirror neurons in debates about the nature of mind-reading supports the positions of the 'neuroskeptics.' Indeed, respecting the linking condition of reverse inference is problematic for virtually any LRI. Nonetheless, we tried to show how PRI provides theoretical tools for overcoming some of these problems. To be sure, this and other new neuroscientific techniques raise a new host of challenges. Still, old adagios such as 'lack of selectivity,' 'excessive functional locationism,' and charges of implementing a 'new phrenology' are not among them. As the field of neuroscience progresses, it is crucial that philosophers and theorists in general do not merely focus on traditional technologies and familiar objections, but also turn their attention to newer and potentially more powerful pattern-decoding techniques, such as PRI. In an attempt to fuel further discussion and constructive debate, in this essay, we suggested a significant distinction between 'two kinds of reverse inference.'

References

Coltheart, M. (2006a). 'Perhaps Functional Neuroimaging Has Not Told Us Anything About the Mind (So Far).' *Cortex 42*, 422–27.

Coltheart, M. (2006b). 'What Has Functional Neuroimaging Told Us About the Mind (So Far).' *Cortex 42*, 323–31.

Coltheart, M. (2013). 'How Can Functional Neuroimaging Inform Cognitive Theories?' *Perspectives on Psychological Science 8*(1), 98–103.

Del Pinal, G. and M. J. Nathan (2013). 'There and Up Again: On the Uses and Misuses of Neuroimaging in Psychology.' *Cognitive Neuropsychology 30*(4), 233–52.

Fodor, J. (1974). 'Special Sciences (Or: The Disunity of Science as a Working Hypothesis).' *Synthese 28*, 97–115.

Fodor, J. (1992). 'A Theory of the Childs Theory of Mind.' *Cognition 44*, 283–296.

Fodor, J. A. (1997). 'Special Sciences: Still Autonomous After All these Years.' *Nus 31*, 149–63.

Fodor, J. A. (1999). 'Let Your Brain Alone.' *London Review of Books 21*.

Fogassi, L., P. F. Ferrari, B. Gesierich, S. Rozzi, F. Chersi, and G. Rizzolatti (2005).) 'Parietal Lobe: From Action Organization to Intention Understanding.' *Science 308*, 662–67.

Formisano, E., F. De Martino, M. Bonte, and R. Goebel (2008). "Who' is saying 'What'? Brain-Based Decoding of Human Voice and Speech.' *Science 322*, 970–73.

Gallese, V. and A. Goldman (1998).). 'Mirror Neurons and the Simulation Theory of Mind-Reading.' *Trends in Cognitive Sciences 2*(12), 493–501.

Gopnik, A. and H. M. Wellman (1992). 'Why the Childs Theory of Mind Really is a Theory.' *Mind and Language 7*(1-2), 145–71.

Harrison, S. A. and F. Tong (2009). 'Decoding Reveals the Contents of Visual Working Memory in Early Visual Areas.' *Nature 458*, 632–35.

Haxby, J. V., A. C. Connolly, and J. Swaroop Guntupalli (2014). 'Decoding Representational Spaces Using Multivariate Pattern Analysis.' *Annual Review of Neuroscience 37*, 435–56.

Hickok, G. (2014). *The Myth of Mirror Neurons*. New York: Norton.

Hutzler, F. (2014). 'Reverse Inference Is Not a Fallacy Per Se: Cognitive Processes Can Be Inferred from Functional Imaging Data.' *Neuroimage 84*, 1061–69.

Iacoboni, M. (2009). 'The Problem of Other Minds Is Not a Problem: Mirror Neurons and Intersubjectivity.' In J. A. Pineda (Ed.), *Mirror Neuron Systems*, pp. 121–33. New York: Humana Press.

Iacoboni, M., I. Molnar-Szakacs, V. Gallese, G. Buccino, J. C. Mazziotta, and G. Rizzolatti (2005). 'Grasping the Intentions of Others with Ones Own Mirror Neuron System.' *PLOS Biology 3*, e79.

Kilner, J. M. and R. N. Lemon (2013). 'What We Know Currently About Mirror Neurons.' *Current Biology 23*, R1057–1062.

Kriegeskorte, N. (2011). 'Pattern-Information Analysis: From Stimulus Decoding to Computational-Model Testing.' *Neuroimage 56*, 411–21.

Lindquist, K. A., T. D. Wager, H. Kober, E. Bliss-Moreau, and L. F. Barrett (2012). 'The Brain Basis of Emotion: A Meta-Analytic Review.' *Behavioral and Brain Sciences 35*, 121–202.

Machery, E. (2014). 'In Defense of Reverse Inference.' *British Journal for the Philosophy of Science 65*(2), 251–67.

Miller, G. (2008). 'Growing Pains for fMRI.' *Science 320*, 1412–1414.

Nathan, M. J. and G. Del Pinal (2015). 'Mapping the Mind: Bridge Laws and the Psycho-Neural Interface.' *Synthese Online First*.

Norman, K., J. Quamme, and E. Newman (2009). 'Multivariate Methods for Tracking Cognitive States.' In K. Rosler, C. Ranganath, B. Roder, and R. Kluwe (Eds.), *Neuroimaging of Human Memory: Linking Cognitive Processes to Neural Systems*. Oxford University Press.

Oosterhof, N. N., S. P. Tipper, and P. E. Downing (2013). 'Crossmodal and Action-Specific: Neuroimaging the Human Mirror Neuron System.' *Trends in Cognitive Sciences 17*(7), 311–18.

Poldrack, R. A. (2008). 'The Role of fMRI in Cognitive Neuroscience: Where Do We Stand?' *Current Opinion in Neurobiology 18*, 223–27.

Poldrack, R. A. (2011). 'Inferring Mental States from Neuroimaging data: From Reverse Inferences to Large-Scale Decoding.' *Neuron 72*(692-97).

Poldrack, R. A., J. A. Mumford, and T. E. Nichols (2011). *Handbook of functional MRI data analysis*. Cambridge, UK: Cambridge University Press.

Rizzolatti, G. and L. Craighero (2004). 'The Mirror-Neuron System.' *Annual Review of Neuroscience 27*, 169–92.

Rizzolatti, G., L. Fadiga, V. Gallese, and L. Fogassi (1996). 'Premotor Cortex and the Recognition of Motor Actions.' *Cognitive Brain Research 3*, 131–41.

Rizzolatti, G., L. Fogassi, and V. Gallese (2001). 'Neurophysiological Mechanisms Underlying the Understanding and Imitation of Action.' *Nature Reviews Neuroscience 2*, 661–70.

Rizzolatti, G., L. Fogassi, and V. Gallese (2009). 'The Mirror Neuron System: A Motor-Based Mechanism for Action and Intention Understanding.' In M. Gazzaniga (Ed.), *The Cognitive Neurosciences*, Volume IV, Chapter 43, pp. 625–40. MIT Press.

Rizzolatti, G. and C. Sinigaglia (2010). 'The Functional Role of the Parieto-Frontal Mirror Circuit: Interpretations and Misinterpretations.' *Nature Reviews Neuroscience 11*, 125–46.

Satel, S. and S. Lilienfeld (2013). *Brainwashed: The Seductive Appeal of Mindless Neuroscience*. New York: Basic Books.

Shallice, T. and R. Cooper (2011). *The Organization of Mind*. Oxford University Press.

Spaulding, S. (2012). 'Mirror Neurons Are Not Evidence for Simulation Theory.' *Synthese 189*, 515–534.

Tong, F. and M. S. Pratte (2012). 'Decoding Patterns of Human Brain Activity.' *Annual Review of Psychology 63*, 438–509.

Tressoldi, P. E., F. Sella, M. Coltheart, and C. Umilta (2012). 'Using Neuroimaging to Test Theories of Cognition: A Selective Survey of Studies from 2007 to 2011 as a Contribution to the Decade of the Mind Initiative.' *Cortex 48*, 1247–1250.

Umilta, M. A., E. Kohler, V. Gallese, L. Fogassi, L. Fadiga, and C. Keyser (2001). 'I Know What You Are Doing. A Neurophysiological Study.' *Neuron 31*(155-65).