The Future of Reverse Inference: Lessons from Mirror Neurons and Mindreading

Guillermo Del Pinal

Center for General Linguistics (ZAS), Berlin

August 19, 2015

Abstract

Cognitive Neuroscientists regularly 'reverse infer' cognitive processes from patterns or locations of neural activation. This article examines the limits and prospects of this controversial technique. To frame the discussion, we consider a famous reverse inference involving mirror neurons and theories of low-level mindreading. The basic claim is well-known: the activity profiles of mirror neurons are taken to support simulationist accounts of how we understand the actions of others. In the first part of this article, I show that the evidence provided by mirror neurons does not favor any of the competing cognitive-level theories of mindreading. Drawing on those results, I defend two claims about reverse inference. Firstly, the most widely used and discussed subclass of reverse inference is inherently problematic. What is distinctive about this familiar class is that the *locations* of neural activation play a key role in the inferences to cognitive processes. Secondly, the problems faced by location-based inferences do not apply to inferences based on pattern decoding techniques. These techniques have been almost completely ignored in discussions of reverse inference in Philosophy of Mind and Science. This should change. Pattern decoding methods overcome some of the most resilient objections which have been raised against both location-based reverse inference in particular and Cognitive Neuroscience in general.

Keywords: reverse inference, mirror neurons, mindreading, simulation theory, theory theory, Cognitive Neuroscience, multivariate pattern analysis, machine learning
Word count: 8,145

1 Introduction

Cognitive Neuroscience is based on the sensible idea that we can understand the mind by examining the workings of the brain. This approach has advanced our theories of relatively modular systems such as vision and touch, and promises to resolve fundamental disagreements in domains of higher-cognition such as language processing and decision making. To achieve the latter aim, one of the most common yet controversial techniques is 'reverse inference', roughly, the practice of inferring the engagement of cognitive processes from patterns or locations of neural activation.[1] Since competing theories often make different predictions about the cognitive processes underlying some task, reverse inference can, in principle, be used to discriminate between theories. In what follows, we examine the limits and prospects of reverse inference.

To frame the discussion, we analyze a case often presented as a clear success of the practice, the case of mindreading and mirror neurons. 'Mindreading' is the capacity to determine and predict the mental states and behavior of others. Explanations of this capacity have resulted in a dispute between two families of cognitive theories—the 'Theory theory' (TT) and the 'Simulation theory' (ST):

(TT) Mindreading is based on a science-like folk theory of mind (Gopnik and Wellman, 1992; Fodor, 1992). This theory includes (law-like) generalizations over symbolic representations of categories of: (i) observable inputs and mental states, (ii) mental states and other mental states, (iii) mental states and observable outputs.[2]

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

In emphasizing the symbolic and abstract nature of higher-cognition, TT follows the 'classical cognitivist' tradition of Descartes and Kant, and more recently, of Chomsky and Fodor. In contrast, ST—in the cases we will focus on—is closer to the idea that the vehicles of cognition have a sensory-motor format, an idea that can be traced to the empiricist tradition of Locke and Hume.[3]

[1] For brevity, I will use the term 'cognitive process' to refer to psychological processes such as recognizing an item as old *and* to psychological states and representations such as being in fear or thinking about a red cup.

[2] As is well known, there are nativist and non-nativist versions of how the 'folk theory' posited by TT is acquired. In addition, there are different accounts of what is meant by having a 'folk psychological theory'. For our purposes, nothing hangs on these choices.

[3] There are caveats and qualifications looming in the background here. More on this later. For now, we take TT and ST in their most popular versions. Regarding the choice of classical cognitivism or empiricism, one can of course favor global or local varieties. One could pair ST with a modal theory when dealing with certain subclasses of mindreading, and with an amodal theory when dealing with others. For a clear exposition of the options, see Goldman (2012).

The idea that neuroscientific data can resolve this fundamental dispute about the nature of mindreading—at least within a certain subclass—gained serious adherents with the discovery of mirror neurons (Rizzolatti et al., 1996, 2001). These (mostly) premotor neurons have the special property that they are selectively activated when subjects observe and when they execute an intentional motor act such as grabbing a cup.[4] In what have become some of the most cited experimental and theoretical papers in the field, prominent neuroscientists and philosophers argue that mirror neurons provide strong—some think decisive—evidence for ST (Rizzolatti et al., 1996; Gallese and Goldman, 1998a; Iacoboni et al., 2005). If correct, this would be impressive. The discovery of mirror neurons would not only count as a victory of ST over TT. It would also count as a triumph for Cognitive Neuroscience, a clear example in which a debate between two fundamentally different approaches to cognition was given a *direct* neuroscientific resolution (Ramachandran, 2000; Rizzolatti and Craighero, 2004; Iacoboni, 2008).

Despite the enthusiasm for mirror neurons and ST, I will show that, for cognitive-level theories of mindreading, the evidence provided by mirror neurons is demonstrably irrelevant—at least, it does not favor ST over TT (or vice-verse), in any of their versions.[5] I will argue that the problem with the inference from mirror neurons is an instance—a very informative one, as we shall see—of a general problem faced by one type of reverse inference, namely, those based on *locations* of neural activation. Location-based reverse inferences are still widely used in Cognitive Neuroscience and Philosophy of Mind. In the final part of this essay, I present and defend a different type of reverse inference based on pattern decoding techniques. I argue that these decoding techniques—increasingly important in Neuroscience yet almost completely ignored in Philosophy—overcome some of the oldest and most resilient objections which have been raised against both location-based reverse inference in particular and Cognitive Neuroscience in general.

2 Mindreading and Mirror Neurons

To discuss reverse inference, examining the case of low-level mindreading and mirror neurons is especially useful. Firstly, the relevant neural-level data is quite clear and has been widely replicated. Secondly, the data was obtained using single cell recordings. This allows us to focus on reverse inference as such without dealing with problems due to more controversial data gathering tools such as fMRI. Thirdly, the debate between theories of mindreading is of

[4] In what follows, I use the term 'mirror neurons' to refer to what some authors call 'action mirror neurons'. Some authors also use the term in a wider sense. Goldman (2009), for example, defines 'mirror neurons' as the class of neurons that 'discharge when an individual undergoes a mental or cognitive event *endogenously* and when it observes a sign that another individual undergoes or is about to undergo the same type of mental event'. I focus on action mirror neurons because they are by far the best studied in the class.

[5] Other authors have objected to the use of mirror neurons to support ST. I especially benefitted from Spaulding (2012), Hickok (2014), and Caramazza et al. (2014).

intrinsic philosophical interest. The views subjected to reverse inference are not toy examples; they are substantial positions held by prominent authors.[6] Other advantages of this example will emerge in the course of the discussion. Here is the plan for this section: §2.1 introduces the mindreading task and spells out the predictions of TT and ST; §2.2 presents the neural-level results; §2.3 presents the reverse inference used to support ST and discusses in detail why it fails.

2.1 Mindreading task: predictions of TT and ST

'Mindreading' involves various capacities. We focus on the subclass which provided the most interesting neuroscientific results: the capacity to understand 'basic level motor acts', i.e., simple motor acts such as grabbing a cup. These motor acts can be productively combined to form complex acts with immediate and distal goals, such as grabbing a cup to drink from it. ST theorists argue that mirror neuron data obtained from cases in which subjects observe and when they execute basic level motor acts decisively favors ST over TT (Gallese and Goldman, 1998a; Iacobani, 2009; Rizzolatti and Sinigaglia, 2010). To establish this claim, ST theorists would have to show that ST predicts the neural-level data *and* that TT does not, at least not as strongly. ST theorist try to establish the former claim, but they usually do not carefully consider the predictions of TT. In contrast, we begin by spelling out the predictions—at the cognitive and neural levels—of each of these theories.[7]

The key data about mirror neurons was obtained using variations of the following basic experimental contrast:[8]

(execution) Subjects perform a basic level motor act, e.g., grabbing a cup or grabbing a cookie to eat it.

(observation) Subjects then observe another agent performing the *same* type of motor act.

In what ways are the cognitive process underlying these conditions similar, and in what ways different, according to TT and ST? Answering this question will help us specify the neural-level predictions of each theory.

[6] This might seem trivial, but as Coltheart (2006b) points out, methodological discussions that try to show that reverse inference is useful for Cognitive Neuroscience often illustrate the practice with cognitive-level theories that are caricatures of real theories. Coltheart (2006a, 2013) challenged neuroscientists to provide examples of reverse inference which adjudicate between real theories held by real theorists. The results were not pretty, to say the least.

[7] In what follows, I present TT and ST in their simplest and most common versions, focusing on their respective accounts of basic level motor acts. Once the main argument is presented, we will be in a position to see that it also applies to other versions of these theories. In particular, here TT is paired with an amodal theory of concepts, and ST with a modality-specific theory. This is how most proponents of each theory present and endorse them, but these pairings are arguably optional: we can formulate a modal TT and an amodal ST. For our purposes, this does not matter. We will argue that mirror neurons do not favor any of the theories as usually construed. It will be clear, once we present the argument, that the conclusion can be extended to the other versions.

[8] For recent reviews of the main experiments and results see Rizzolatti et al. (2009) and Kilner and Lemon (2013).

According to TT, subjects posses concepts—conceived as amodal symbolic representations—for motor actions, e.g., GRAB. Since these are intentional concepts, we can represent them as having slots for an agent x and the object y of the action event e, e.g., '$\text{AGENT}(e, x) \wedge \text{GOAL}(e, \text{GRAB}(x, y))$'. These concepts can be combined with other concepts to represent more complex basic level motor acts, such as when we represent x as grabbing a cup to drink from it. These concepts form part of folk-psychological generalizations such as that subjects that grab cups to drink are often thirsty. In addition, although these concepts are amodal, they interface with sensory and motor processes. On the sensory side, certain perceptual cues can be used to apply these concepts. On the motor side, these concepts can be used to form intentions to do certain things. So when we form the intention to GRAB THAT RED CUP, we can usually execute the corresponding motor act, so that in this case we will indeed grab and not bite the red cup.[9]

Here's the crucial point: according to TT, the concept GRAB is used when we categorize the grabbing actions of others, including complex ones, when we imagine actions of grabbing, and when we form the intention to grab something for whatever reason. So TT entails, at the cognitive level, that when a grabbing act is either perceived as such or intentionally executed, the concept GRAB is tokened.[10] In other words, TT predicts that both **execution** and **observation** involve tokens of GRAB, or in the more complex case, tokens of GRAB CUP $\wedge$ TO DRINK. What does this entail at the neural level? That the neural implementation (e.g., the neural pattern of activation) that codes for tokens of the concept GRAB should be instantiated in cases of both **execution** and **observation** of grabbing actions. Of course, TT also predicts that there should be some differences between the conditions: most obviously, the representation of the agent of the grabbing act is different.

Consider next the predictions made by ST. On this view, simulation is the default process we use to mindread others, including understanding their basic level motor acts. When subjects perceive others as grabbing a cup, they can simulate this action as if they were executing it. This simulation allows them to retrodictively determine what intentions they would have when doing that act in those conditions.[11] The emphasis in this case is on the simulation of a motor

[9] These interface conditions are sometimes call 'legibility constraints'—the idea being that, to be usable at all, a conceptual system must be legible at its interfaces. Concepts, even if amodal, must be legible at the input-output interfaces: perceptual input can lead to tokenings of certain concepts, and tokenings of certain concepts can be translated into the appropriate motor commands. In addition, some authors argue that concepts, even if amodal, have to be reformatted to be usable by the Faculty of Language (Pietroski, 2010).

[10] The need for cognitive control and monitoring also means that, on a TT model, we often token concepts for motor acts, not only when we form the intention to execute them, but also as we carry out and monitor those very acts. This is perhaps easier to see in more complex motor acts, such as baking a cake.

[11] There are different ways of spelling out the details of the action simulation process. According to one option, closely associated with Rizzolatti et al. (2001), the process is rather direct, in the sense that the observed motor action is directly matched with a corresponding motor representation of that action, which initiates the process of understanding that action from the inside. In a more complex model, presented initially by Gallese and Goldman

processes. The key thing to note, for our purposes, is that **observation** tokens a processes that is a subcomponent of, or is at least structurally analogous to, the process involved in the actual execution of the same type of action. So ST predicts, at the cognitive level, that a subset of the process underlying **execution** is also tokened in **observation**, i.e., when understanding the basic level motor acts of others. What does this entail at the neural level? That the neural implementation of the part of the simulation process that is shared with the execution motor process should be instantiated in both **execution** and **observation**.

At this point, we can see that distinguishing between TT and ST, in the experimental contrast captured by **execution** and **observation**, is not as straightforward as is commonly thought. TT and ST both predict that there is a key cognitive component present in both experimental conditions. According to TT this is the tokening of particular action concepts, and according to ST this is the tokening of particular subsets of motor processes. Still, if there is way to distinguish, at the neural level, tokens of action concepts from tokens of motor action simulation processes, then neural data could, in principle, distinguish between these theories.

2.2 Neural-level results and Mirror Neurons

We said that the neural-level results were obtained by comparing two basic conditions: in **execution** subjects execute a basic level motor act, and in **observation** subjects observe full or partial evidence that others are executing the same type of basic level motor act. The key results were obtained in studies using single cell recordings of macaque monkeys (for overviews, see Rizzolatti et al. (2009) and Kilner and Lemon (2013)), but fMRI studies suggest that they also apply to humans (Rizzolatti and Craighero, 2004). At any rate, this is commonly assumed by researchers who favor ST, and here I will accept this reasonable extension. The key neural-level results are as follows:

(MN pattern) There is a group of neurons, say '$n_1 \dots n_{10}$', that are selectively activated in **execution** and **observation**.

(MN location) Neurons $n_1 \dots n_{10}$ (i.e., neurons with that activation profile) are localized in the premotor cortex.

(1998a), subjects 'generate' a candidate action goal for the observed motor action. They then simulate the observed action as if it was their own; if the action matches the goal, then that is the goal of the other's action. In case of mismatch, the process can be repeated. For our purposes, we can remain neutral between these models; but it seems to me that the second model is rather implausible as a default. On that model, there must be a mechanism which generates the possible action goal on the basis of the observation, i.e., of the perpetual input. However, this mechanism must be at least fairly reliable: since, in any given case, there are innumerable possible candidate action goals, and the assignment is usually quite fast, the mechanism which generates the candidate action goal has to use the perceptual cues reliably. If that is the case, the extra step of 'confirming' that the generated goal is the goal of the observed action would be, in general, both costly and superfluous.

Note two important features of these results. First, to say that some neurons are 'selectively activated' in certain tasks is to say that they are *not* differentially activated in other relevant tasks. (a) Mirror neurons are not differentially activated by acts that look similar to but are not motor acts: e.g., by similar motions of a hand that are not instances of a grabbing action (Rizzolatti et al., 1996). (b) Mirror neurons are activated by instances of the same type of basic level motor act even if the cues for the act vary in modality: e.g., they are activated not only when subjects see the full action of grabbing, but also when they see only part of it or even if they only hear evidence of the act (Umilta et al., 2001). (c) The mirror neurons activated in **execution** and **observation** of a basic level motor act such as grabbing a cup, are also activated when that act is part of a more complex one such as grabbing a cup to drink. Specifically, if you compare a set of **execution** and **observation** conditions involving a complex motor act such as grabbing a cup to drink with a set involving another complex act such as grabbing a cup to clean, some mirror neurons fire in both sets of conditions (both sets involve grabbing acts) and some only in one or the other set (for drinking vs for cleaning) (Fogassi et al., 2005; Iacoboni et al., 2005).

Secondly, I presented the result in terms of two components, the 'MN pattern' and the 'MN location'. Experimenters and theorists who appeal to mirror neurons to defend ST rarely separate these two aspects of the results. Ultimately, most defenders of ST assume that the inference proceeds from the MN location (which, in any case, is richer in that it includes the MN pattern). However, discussing the MN pattern and the MN location separately will allow us to determine how much weight, in the inference to ST, is carried by the location of mirror neurons. At the same time, it is important to point out that Cognitive Neuroscientists employ both types of inferences, i.e., inferences to cognitive process from particular *patterns* of neural activation, and inferences from particular *locations* of neural activation.[12] Neither of these types of reverse inferences, however, should be confused with the pattern-*decoding* techniques which I introduce and defend in §3.

2.3 Reverse inference from Mirror Neurons

In 'reverse inference' the engagement of a cognitive process is inferred from neural data, usually consisting of particular patterns or locations of neural activation in a set of tasks. We can model reverse inference in Bayesian terms, i.e., as the conditional probability that a particular cognitive process is engaged given a set of tasks and patterns or locations of neural activation (Poldrack, 2006).[13] That conditional probability depends on the probability of the neural

[12] Examples of purely pattern-based reverse inference are found in EEG-based studies. For example, several studies in neurolinguistics use the N400 event-related potential as a measure of semantic expectations. Since theories often make different predictions about the degree to which some incoming linguistic item is expected, these measure can be used in reverse inferences. In general, these inferences are made without considering the neural location which causes the N400.

[13] For details of how to implement Bayesian accounts of reverse inference see Hutzler (2013) and Del Pinal and Nathan (2013). For a likelihoodist alternative see Machery (2013). We

activation in the task given the hypothesized cognitive process. Since we are trying to determine whether the mirror neuron results provide any reason to select ST over TT, we need to consider these values for each hypotheses: i.e., we need to determine how likely the neural results are given each of ST and TT. What I will now show is that these hypotheses are indistinguishable in the following sense: both (weakly) predict the MN pattern, and both are merely compatible with the MN location.

Consider first the MN pattern, namely, the result that in both **execution** and **observation** a particular set of neurons selectively fire at the same rate. As we showed above, TT predicts that both conditions engage tokens of the same basic level motor act concept. This means that TT predicts that there should be a uniform neural pattern in both conditions, namely the pattern that codes for these tokens of the same concept. In our example, the relevant neural pattern is precisely the uniform and selective firing rate of $n_1 \dots n_{10}$ in both conditions. ST makes essentially the same neural-level prediction, but in this case the predicted uniformity is due to partially overlapping motor processes being engaged in both conditions.[14] To be clear, what both theories predict is that there should be a uniform neural pattern in both conditions, but neither makes any predictions regarding the particular fine-grained structure of this pattern. In other words, ST theorists do not provide any argument or relevant information about the MN pattern which might allow us to determine whether a particular neural firing *pattern*—e.g., that neurons $n_1 \dots n_{10}$ have the firing rate they do—codes for, say, an amodal representation of an action as opposed to a motor simulation (sub)process. So we cannot determine whether the specific firing rate represented by the MN pattern codes for the representations posited by TT or for the simulation process of ST. We come back to this issue in §3.3-§3.4, where we discuss the advantages of pattern decoding techniques.

Consider next the MN location, i.e., the result that in both **execution** and **observation** the set of neurons that selectively fire at the same rate are found in the premotor cortex. Even if an ST theorist accepts the argument just presented against the relevance of the MN pattern per se, s/he would insist that what really favors ST over TT is the MN location, i.e., that mirror neurons are found in the premotor cortex (Rizzolatti and Sinigaglia, 2010).[15]

The MN location might seem especially relevant due to a key difference between TT and ST. According to TT, what **execution** and **observation** share

return to these issues in §3.1.

[14] Strictly, TT and ST make this prediction only on the assumption that tokens of the same cognitive process should have uniform neural implementations, at some level of neural description. Although in principle one can challenge it (e.g., based on radical forms multiple realizability), this uniformity assumption is a fundamental presupposition of Cognitive Neuroscience. We take it for granted in this discussion.

[15] Other considerations also suggest that ST theorists are impressed by the MN location. As we mentioned in §2.2, mirror neurons fire across modalities. From that pattern, one might be tempted to infer that they are amodal. But ST theorists resist this move, and instead propose that since the mirror neurons are in the premotor cortex, they should be understood as translating various modalities to the motor modality, as opposed to translating various modalities to a non-modal abstract representation.

is the tokening of the same basic level action concept, which is taken as an abstract representation that cannot be reduced to a set of sensory or motor processes. In contrast, ST says that what **execution** and **observation** share is a partial overlap of the same type of motor process, in one case as part of an actual motor act, and in the other as part of a simulation. At this level of informal description, the MN location might seem decisive; for as ST theorists emphasize, most neurons in the premotor cortex are known to be 'involved' in motor acts. This reasoning suggests to ST theorists that it is much more plausible to assume that mirror neurons implement subsets of motor plans than relatively abstract motor action concepts.

Despite the intuitive appeal of the previous argument, ST does not predict the MN location in a stronger sense than TT does. Strictly speaking, both theories are merely compatible with the MN location. The problem with the argument offered by ST theorists has two sources, which I elaborate below. First, they do not consider carefully what TT entails regarding the possible encoding locations for tokens of motor act concepts. Secondly, they assume that ST makes stronger predictions about the neural locations where simulated computations might be implemented than it actually does.

Consider the relation between TT and the MN location. We have seen that TT entails that when planning a motor act such as to grab a cup, agents form an intention that tokens the concept GRAB CUP. This intention then interfaces with and serves as an instruction for 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, conjoined with a representation of a different agent. The question is this: does TT predict that tokens of concepts for basic level motor acts such as GRAB be encoded or implemented in the premotor areas? It is easy to see that TT is at least compatible with the MN location: if the safest assumption about premotor areas is that they are 'involved' in motor actions, then these are very plausible candidate locations for the interface between tokens of motor intentions—which on this view token basic level motor act concepts—and the operations involved in motor executions.[16]

[16] To undermine this TT friendly account of the MN location, one could bring up another bit of evidence that ST theorists often mention (Gallese and Goldman, 1998b). When the neurons in (roughly) the MN location are directly activated—e.g., by using transcranial magnetic stimulation (TMS)—there is a motor evoked potential (MEP) in some of the muscles that would typically be used to execute the act that corresponds to the observed basic level motor act. However, this results can be easily explained by TT theorists, since they plausibly result from priming. Just as any other instance of neural activation, tokening an action concept such as GRAB CUP activates through frequent association other networks. Since motor intentions are frequently followed by motor acts, which token the relevant action concepts, we expect that merely tokening action concepts primes the networks underling the corresponding motor acts. That associative priming is a basic feature of neural computation is hardly in doubt; so invoking it to explain the relation between action observation and motor evoked potentials is not an ad hoc move. In any case, it is arguable that ST also needs to invoke priming to explain the MEP results: after all, simulating a motor act is not the same as executing it. Furthermore, there is strong evidence in support of the priming account of the MEP results. Catmur et al. (2007) present the relevant experiments. They first obtained the standard MEP result by showing that when subjects watched a video of index finger abduction, the MEPs were greater in the subjects own index finger, and when they watched a video of little

There is a general point here worth emphasizing, esp., in the context of debates between amodal and 'embodied' theories of concepts. Currently, nothing we know about the nature of neural computation and representation prevents us from holding that amodal 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.[17] In particular, a TT theorist can hold, 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 clear, this topological closeness between locations of concept encodings/tokenings and their corresponding input and output interfaces is not strongly predicted by TT. For we can imagine and cannot currently dismiss other 'hardware' solutions to processes such as extracting conceptual categories from sensory and motor modalities, applying concept tokens to sensory inputs, and using concepts to form motor intentions that can interface with motor processes to execute actions. Still, the crucial point is that TT does not predict that there should be any 'distance' between the 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. In short, although TT does not strongly predict the MN location, it is certainly not undermined by it either.

This conclusion might not seem completely bad news for ST theorists. However, whether the MN location favors ST over TT depends on whether ST predicts that motor simulation should be implemented in the premotor cortex *in a stronger sense* than TT predicts that the premotor cortex is the interface

finger abduction, the MEPs were greater in their little finger. The experimenters then trained subjects to move their fingers in the opposite manner: i.e., to move the little finger if they watched the video of the index finger movement, and to move the index finger if they watched the video of the little finger. After training, and in a condition when subjects only watch and categorize, MEPs were greater in the little finger when the index finger video was observed, and vice-versa. The MEPs can be best explained as a result of priming/association: for presumably the categorization of the observed action did not change when the MEP pattern was reversed.

[17] Indeed, suppose you build a neural network model in which input from nodes representing auditory modalities or visual modalities activates a node representing an amodal concept (e.g., specific color modality nodes activate the node representing the abstract concept 'COLORED'). This would be a network in which the modal and amodal nodes interface. An easy way to do this is to put the nodes next to each other (i.e., topologically close). Similarly, you can build a network in which input from various modalities can activate a node representing an amodal concept (e.g., either vision nodes or auditory nodes active some abstract concept), and the node representing the amodal concept could be close to some or all of the modal input nodes. Compared to a model in which various modalities are translated into one modality (e.g., in which various modalities are translated into motor nodes) the network with the amodal node would only require one additional level. This additional level could easily be implemented in roughly the same location. If you add considerations of 'cost' of long distance communication (admittedly somewhat vague), then the idea that domains that have a functional interfaces should be topologically close is almost inevitable. Indeed, that our semantic/conceptual knowledge is organized such that visual concepts are encoded in areas 'close' to visual areas, auditory concepts in areas 'close' to auditory areas, and so on, is strongly suggested by investigations of semantic memory and certain neurological disorders such as semantic dementia (Shallice and Cooper, 2011).

between motor intentions and motor execution processes. Now, ST is at least compatible with motor simulations being implemented in the premotor cortex. However, does ST strongly require that? Suppose that the results in **execution** and **observation** showed instead that the MN pattern (mutatis mutandis) was localized in premotor areas in the former task but in the prefrontal cortex in the latter. Add that the prefrontal cortex is commonly thought to process 'higher' cognitive abilities such as deliberate rule-following, goal oriented activity, and other aspect of executive control and decision making (Miller and Cohen, 2001). Would that result have directly undermined ST? Clearly not. The neural location that implements a simulation process could be different from the location that carries out the actual processes that is simulated for various reasons. Here is one obvious candidate: 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 details of the actual firing pattern, independently of the location.[18]

Admittedly, there is a certain scratch-your-head quality to thinking about the locations where the brain could implement motor action concepts and simulation processes. There is a reason for this: we do not know enough about how cognitive representations and operations are extracted, encoded and tokened at the neural level to be able to say something that informatively limits, given the demands of this debate, the sorts of neural locations that could perform the categorization operations of TT or the simulation processes of ST (we come back to this in §3.3). The strongest thing we can say, at this point, is that both TT and ST are compatible with the MN location. To be clear, I am not denying that we know *something* about the function of *premotor* neurons, and by extension, of premotor *mirror* neurons. We know—and at any rate we are assuming—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, in itself, help us select which of the competing operations are actually used.[19]

[18] Of course, it could be determined by details of the firing pattern only if the patterns is sufficiently selective for the simulated motor process. Note that this sort of reasoning is not unusual. This is the way in which experiments often try to show that, although the perirhinial 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.

[19] An additional consideration often used by ST theorists appeals to deficit patterns caused by certain neurodegenerative deceases 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. Space prevents me from discussing these issues in detail. For a careful discussion of these cases which supports the view presented here, see Hickok (2014).

3 Limits and Prospects of Reverse Inference

We have seen that both TT and ST predict the MN pattern. We have also seen that adding the MN location—namely, that the MN pattern is localized in the premotor cortex—does not help discriminate between these theories. We can gain valuable lessons from these results. To do so, I begin by spelling out a key condition for the proper use of reverse inference. I then argue—by drawing on the mindreading example and some famous fMRI studies—that this condition is extremely hard to respect for any study that infers cognitive process from *locations* of neural activation in a task ('L-reverse inference'). Crucially, most studies that try to use neural data to support psychological theories of higher-cognition still use L-reverse inference. This might seem to leave us with a rather bleak picture of the usefulness of Cognitive Neuroscience for the study of cognition. However, in §3.3-§3.4 I present an increasingly influential technique—multivariate pattern analysis—that can be used to support a type of reverse inference that directly overcomes the problems faced by L-reverse inferences.

3.1 Linking Condition and Reverse Inference

Suppose that C and D are two competing cognitive theories of the processes underlying some task t. C says that cognitive process c is engaged, D says that d is engaged, and $c \neq d$. Let n stand for a differential pattern of neural activation in some location, and assume that C is the supported theory:

- A reverse inference from the presence of n in task t to the engagement of c in t depends on other studies which establish a link between n and c. Call these background studies, 'linking studies'.
- In the linking studies that associate n and c experimenters have to *assume* that c is engaged in a set of tasks t^* that activate n. The experimenters must be confident that c is engaged in t^*.
- If those conditions are satisfied, experimenters can infer c from observing n in t. Here reverse inference is used because experimenters do not know whether c is engaged in t.

Linking studies are crucial for reverse inference. At least two things could go wrong:

(P1) C and D could differ not only in their predictions for task t, but also in their cognitive-level interpretation of t^*. If so, the evidence used in the linking studies that associate c and n is problematic: it ignores at least one of the competing theories, in this case D, and is therefore biased against it.

(P2) Even though the linking study of c and n was properly conducted, the region where n is located might also be known to implement other cognitive processes. The less selective that region is, the less confident we can be in the reverse inference from n to c.

Problems P1 and P2 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.

To minimize the possibility of violating the linking condition, properly conducted studies invoke, as part of their background linking studies, tasks that are relevantly different from those subsequently used to discriminate amongst the competing cognitive hypotheses. Specifically, 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.

3.2 Linking Condition and Location-based Reverse Inference

The linking condition is clearly violated by the way in which ST theorists use the MN location to argue against TT. ST theorists do not appeal—to give a cognitive-level interpretation to the MN location—to any tasks in which it is reasonably uncontroversial that something like simulation processes—or one of its key subcomponents—is engaged. Furthermore, in the tasks they do consider—those variations of **execution** that generate mirror neuron activation—what is under dispute by ST and TT is precisely their fine grained functional interpretation. This is a clear instance of P1. In particular, mere activation in premotor areas does not have a cognitive-level interpretation that is relevant to adjudicate between ST and TT. Why? Because ST and TT provide different cognitive-level accounts of the interface between intentions for and execution of basic level motor acts; hence, they provide different accounts of what precisely 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 either the cognitive-level account of TT or ST.

One might think that the mindreading example is special—maybe the differences between TT and ST are too subtle, at least when applied to observing and executing basic level motor acts, or maybe the fine-grained computational diversity of the premotor cortex is unique. However, most reverse inferences used in studies of higher-cognition violate the linking condition in ways that are even more obvious than the mindreading example.[20] There is a non-accidental reason for this. In studies of higher-cognition, location-based reverse inference—in

[20] Essentially this point is defended by Coltheart (2013, 2006a), where he presents a forceful criticism of various recent attempts to use neuroimaging data, via reverse inference, to advance theories of higher cognition. Coltheart examines various famous studies, and in many cases the problems are basically instances of P1 or P2. Hutzler (2013), Machery (2013) and Del Pinal and Nathan (2013) develop more optimistic accounts of location-based reverse inference, which try to respond to some of the worries raised by Coltheart. For reasons developed below, I do not think any of these responses fully address Coltheart's basic worry.

which the engagement of a cognitive process in a task is inferred from a particular location of neural activation—is still the most common technique used to make neural data bear on competing cognitive-level hypotheses.[21] The difficulty of satisfying the linking condition—i.e., of avoiding P1 and P2—when employing L-reverse inference is that most brain regions of interest are not selective—they are engaged by various cognitive processes, ranging from very distinct to only subtly different. This functional diversity usually undermines the usefulness of particular L-reverse inferences.

Consider a famous study of moral decision making. Using fMRI, Greene et al. (2001) scanned subjects making decisions in two kinds of tasks, all of which involve choosing whether to sacrifice some people to save more. In one set all the choices that would save the majority involve using others directly as a means (*personal cases*), and in the other set the majority can be saved without using others as a means (*impersonal cases*). Greene and colleagues argue that the resulting patterns of activation support theories of moral decision making that posit the involvement of emotions in *personal cases* over theories that assume that such decisions are based on some form of rule application, such as deontological theories. Their reasoning is straightforward: in previous linking studies, the brain regions differentially activated in *personal cases* were associated with processes involving negative emotions. In response to this study, various authors point out that the brain regions used in the L-reverse inference are quite un-selective (Poldrack, 2006; Phelps and Delgado, 2009; Mole and Klein, 2010). For example, the amygdala—one of the brain regions invoked—is associated with negative emotions, but linking studies have also associated it with: odor intensity, sexually arousing stimuli, trust from faces, faces from other races, biological motion, shape contours, and novel stimuli (Phelps, 2006; Lindquist et al., 2012). This lack of selectivity is a property of most brain regions. Indeed, meta-analyses suggest that most of the brain regions commonly used in influential examples of L-reverse inference—e.g., Brocca's region, the insula, and the DLPFC—are extremely un-selective (Yarkoni et al., 2011).

One could argue that the lack of selectivity of most brain regions does not completely undermine the usefulness of L-reverse inference, although it should certainly serve as a general warning against excessive confidence in particular instances. However, the problem is not primarily that activation in a region could signal the engagement of 'some' cognitive process other than the one predicted by the 'supported' theory. Recent models of reverse inference have addressed this problem (Hutzler, 2013; Del Pinal and Nathan, 2013). For example, Machery (2013) presents a likelihoodist account which takes reverse inference as an inherently comparative technique. From this perspective, experimenters can ignore cognitive processes associated with a brain region that are not part of at least one of the competing theories. In the moral decision example, it does not then matter that the amygdala could be processing things like odor intensity and sexually arousing stimuli, since these possibilities are not parts of the

[21] For example, in *The Organization of Mind* by Shallice and Cooper (2011), a recent and highly regarded advanced textbook on the Cognitive Neuroscience of Higher-Cognition, the methodological section is almost completely devoted to L-reverse inference.

hypotheses being tested. The problem, however, is that when regions are so un-selective differential activation could signal—and in actual examples often does—the engagement of cognitive processes directly relevant to the competing and supposedly undermined theories. This seriously limits the usefulness of this technique. For, as illustrated in the mindreading case, important debates often involve disputes about the fine-grained cognitive functions carried out in particular tasks and neural locations.

Going back to the moral psychology example, we can see that some plausible candidates relevant to the competing hypothesis remain. For instance, the amygdala could be processing novel stimuli. It is arguable that, in general, *personal cases* are more novel/weird than *impersonal cases.* We often have to sacrifice someone at a distance, so to speak; but rarely do we have to directly use someone as a means, as in the footbridge trolley case. Another possibility is biological motion, mostly present only in the *personal cases.* Greene et al. (2001) also mention, in support of their conclusion, that there is less activation in the dorsolateral prefrontal cortex (DLPFC) in *personal* compared to *impersonal cases.* They note, as their linking studies, that the DLPFC has been associated with cognitive control, including following rules (Miller and Cohen, 2001). However, theories that posit the rule-guided nature of moral decision making arguably also predict this result. Plausibly, applying consequentialist rules that involve calculation (such as is required in the *impersonal cases*) is more effortful than applying categorical rules (such as the 'do not use someone as a means' presumably used in *personal cases*). This might account for the increased activation pattern. In addition, in a meta-analysis of the base-rate of activation of brain regions of interest commonly used in L-reverse inference, the DLPFC came out as one of the *least* selective regions, being differentially activated in around 20% of all 3,489 studies (Yarkoni et al., 2011).[22]

Crucially, the tension between lack of selectivity and L-reverse inference 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, such as 'negative emotions' or 'following rules'. This problem is illustrated by fMRI based-cases such as the moral psychology example. At a much finer level, the difficulty arises from the computational diversity

[22] One area of study where it is arguable that L-based reverse inference has been used with some success is decision-making. I think that this is the fortunate result of a set of unusual conditions in the field. As Camerer and colleagues (2005) argue in detail, one of the main divisions in current studies of decision making is between models that post purely rational processes, and 'mixed' models that posit the essential involvement of emotions. This division is exemplified in the moral decision example and in several debates in neuroeconomics, such as in competing explanations of the Endowment Effect (Knutson et al., 2008). This clean and extreme division (including the relative lack of overlap in the cognitive mechanisms posited by competing theories of each type), coupled with decent linking studies that map emotion and rule-guided processes to with arguably different brain implementations (Miller and Cohen, 2001; Greene, 2009), allows for the limited use of L-reverse inference. However, as this branch of science progresses and mixed models that incorporate both rational and emotional components become more common—as has been argued is increasingly the case by e.g., Phelps (2009) and Phelps and Delgado (2009)—it will be much more difficult, if not impossible, to use L-reverse inference to discriminate amongst them in neuroimaging studies.

associated with single neurons or groups of neurons in a given region. This problem is perfectly illustrated by the case of mindreading and mirror neurons. 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 much 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. Clearly it is then 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.

3.3 Pattern Decoding and Reverse inference

Meeting the linking condition, we have seen, is extremely hard for L-reverse inference. The source of the problem is the computational diversity of brain regions, which occurs at various levels of description. We now introduce a type of reverse inference that can respect the linking condition. Consider again the mindreading example, and let us ask what sort of comparison, according to the linking condition, we would need to properly test TT and ST. We would need to set up tasks in which the cognitive processes posited by ST (motor act simulation) are uncontroversially engaged, and tasks in which the cognitive processes posited by TT (categorization) are uncontroversially engaged, and in each case map the neural results to the corresponding cognitive-level processes. We could then compare those neural-level results with the results obtained in **observation**, a task in which the identity of the underlying cognitive process is under dispute. By comparing the resulting neural patterns, we could determine whether understanding the basic level motor acts of others is more like a simulation process or more like a straightforward categorization process.

This is just the type of comparison that we can perform with pattern decoding techniques such as multivariate pattern analysis (MVPA).[23] The basic idea is this. Although, as we have seen, mappings between patterns of activity in particular brain regions—going down to single cells—and functions might not allow for selectivity, mappings between multivariate vector patterns of activity

[23] The idea that reverse inference should be based on decoding methods such as MVPA has been pushed by Poldrack (2008, 2011). The following discussion is indebted to Poldrack's insightful work. For overviews of MVPA and related methods see, in addition, Poldrack et al. (2011) and Tong and Pratte (2012).

and functions are, as we will see, much more selective. In other words, vector patterns of neural activity contain detailed information about even fine grained cognitive-level representations and processes. To introduce this technique, consider again the MN pattern, which says that some set of neurons fire at the same rate in **observation** and **execution**. In the canonical studies of mirror neurons by ST theorists, this firing pattern has not been decoded in any serious sense. That is to say, we do not know what sort of process or content is likely encoded by that sort of firing pattern; as we showed above, that simple firing pattern could implement a token concept for a motor act, or a subset of a motor processes, among various other possibilities. MVPA uses tools from machine learning to create statistical machines—called 'classifiers'—which can, often quite accurately, decode the cognitive states or process encoded in particular neural data sets, such as multivoxel patterns obtained using fMRI.

To illustrate MVPA, consider a case from episodic memory research which has interesting parallels with the mindreading debate. We focus on a debate about the cognitive processes underlying our capacity to reliably recognize items as old or new. For concreteness, let us formulate this recognition capacity as follows (assume 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 with item $e \in E$ and has to decide whether she has previously encountered e, s can reliably distinguish between old and new items. Most memory researchers now accept versions of a dual process recognition theory. Two competing explanations, recently advanced, are the following:

(R) 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.*

R and RF use the same basic components to explain recognition decisions, but they posit different interactions between them. According to R, subjects generally use recollection information to decide whether items are old, and only use familiarity when such information is unavailable. In contrast, RF predicts that certain contextual cues will induce subjects to make familiarity-based recognition decisions even if recollection information is available.

To test these theories, 'pattern classifiers' are trained to determine the multivoxel patterns associated with recollection and familiarity processes. Specifically, classifiers are trained in tasks where experimenters can control which cognitive process is engaged, thereby explicitly meeting one of the linking condition 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 are divided into two sets: *recollection blocks* and *familiarity blocks*. In recollection blocks, subjects are instructed to recall specific details of the mental image formed during the study phase, and to only answer 'yes' if they are successful. In contrast, in familiarity blocks subjects are instructed to answer 'yes' if the word is familiar and to ignore any details they might recollect from the study phase. After training, classifiers can determine whether some multi-voxel pattern of neural activation is an instance of recollection or familiarity.

What gives MVPA a big advantage over traditional L-reverse inferences is that the reliability of the classifiers can be established within the experiment. In this example, this can be done by saving a subset of the recollection and familiarity blocks for later testing (so they are not used at the training stage), and then determining the rate at which the classifier correctly categorizes the corresponding neural patterns. This part of the study, in which experimenters can control which process is engaged, establishes the links between recollection, familiarity and their corresponding multi-voxel patterns that can then be used in reverse inference.

Having established the links, one can then test competing hypotheses R and RF in cases where the engagement of the sub-processes cannot be directly controlled. So in a second phase of the study, subjects were scanned while trying to determine whether some word is old or new, while being exposed to previously studied items ('shoe' and 'ball'), unrelated lures ('horse' and 'box'), and previously unstudied switch-plurality lures ('balls'). To test the hypotheses, experimenters then examined the subset of the items for which subjects made correct positive recognition decisions. Note that these are cases where both recollection and familiarity information was available to subjects.

- According to R, the classifier should categorize the corresponding voxel patterns as recollection patterns (since this is the default).
- According to RF, the classification should be more variable, involving—at least in some cases—familiarity patterns.

The MVPA experimental results support RF over R (Norman et al., 2009). When both types of information are available, various contextual cues determine whether subjects use familiarity or recollection as the basis of their recognition decision. That is to say, contextual cues determine whether, according to the classifier, the multi-voxel patterns underlying recognition decisions were more like unambiguous familiarity or like unambiguous recollection patterns.

3.4 Advantages of Pattern Decoding-based Reverse Inference

MVPA has substantial advantages over L-reverse inference. Three immediately stand out for their experimental and philosophical relevance.

1. The reliability of classifiers can be precisely determined within a phase of the experiment.

This feature of MVPA fulfills the linking condition, and allows experimenters to quantify their confidence in particular reverse inferences. This is a substantial advantage over L-reverse inference. As we discussed, successful reverse inferences presuppose successful linking studies. The linking studies required for L-reverse inferences face several difficulties, stemming from the lack of selectivity of most brain regions of interest. For example, when the differences between the competing theories are fine-grained, the linking studies often ignore one of the theories. This is clear in the ST-TT debate, where the functional interpretation of premotor areas (esp. in **execution** tasks) according to TT is basically ignored. In addition, even if the linking studies assumed by a L-reverse inference are, in themselves, non-problematic, we still have to consider other tasks that activate the brain regions of interest and ask whether these might be relevant for the theories we are comparing.[24] The problem is not primarily due to the resolution of the tools commonly used in L-reverse inference, i.e., fMRI. This is partly why the case of mirror neurons is instructive. The single cell recording method reveals quite starkly the fine-grained computational diversity of neurons in the premotor cortex of macaques. As a result of this computational diversity, which goes all the way down to single neuron function, it is often impossible to determine the degree to which an L-reverse inference should increase our confidence in a hypothesis, if at all. In contrast, MVPA can decode cognitive processes from multidimensional vector patterns—rather than regions—of neural activation. In any given case, the reliability of the classifier that underlies a particular reverse inference can be established within the experiment. Of course, the success of particular classifiers depends, among other things, on the type of task, the amount of learning trials, the brain regions used for analysis (if restricted), and the machine learning algorithms (Poldrack et al., 2011). In the recognition memory example, classifier accuracy was around 60 percent. In other tasks—e.g., in tasks that use classifiers to determine the basic level categories of objects—classifier accuracy can be much higher (Haxby et al., 2014).

2. Classifiers do not ‘assume’ that the cognitive states or processes of interest are functionally localized.

The multivoxel patterns used by classifiers can be distributed across the brain. Of course, experimenters can restrict the analysis to particular brain regions, esp. if there is prior reason to think that some brain region is crucial for some task, or if what is being tested is the degree to which a region is responsible

[24] Those who have had to go through meta-analysis of the functions associated with brain regions know how frustrating and confusing this can be. For a discussion of some of the difficulties, see Poldrack (2011). For a serious exploration of all the possible functions of regions commonly used in various decision making studies—which will likely cause you to loose faith in the L-inferences of those studies—see Lindquist et al. (2012).

for executing some task.[25] Whether the restriction is useful is revealed by the performance of the classifier, and is not in any sense assumed. Furthermore, classifiers can also take widely distributed multi-voxel patterns. This is especially useful when comparing complex multi-step processes that likely involve various brain regions. All the examples we have considered here are of this kind, as are most other models of higher-cognitive capacities. Hence MVPA does not fall prey to one of the oldest and most resilient objections against the traditional methods of Cognitive Neuroscience, including L-reverse inference: that it *assumes* a strong and objectionable form of functional locationism. If only for this reason, this method is of immediate philosophical interest. From this perspective, the degree of modularity, specialization, and functional localization of various cognitive processes and brain areas will emerge as a result of studies using MVPA. For example, MVPA studies consistently show that the ventral temporal cortex carries sufficient information for classifiers to reliably distinguish between animate and inanimate objects (Kriegstone 2008). In the case of mindreading, we have seen that the MN pattern and location provide too little information to discriminate between competing theories. 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, the first few 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).

3. The information decoded from multivariate vector patterns does not depend on previous informal assumptions about the coarse 'function' of a given brain region of interest.

We have seen that assuming that some brain region of interest has a given coarse function does not help when using L-reverse inferences to try to adjudicate between theories with fine-grained differences. One example is the ST-TT debate, where the crucial distinction is between a motor simulation process and the interface between an abstract representation of a motor goal and a motor execution process. Another example is the debate between R and RF, where the key distinction turns on the dynamics—not the components—of recognition decisions. Once again, single cell recordings, such as those used in studies of mirror neurons in monkeys, are extremely instructive, for they highlight the

[25] To illustrate, many neuroscientists think that the hippocampus is the locus of spatial processing, and that the perirhinal cortex is involved in processing single item information. To test these hypotheses, classifiers can be trained to decode spatial (or item) information from each of these regions, and the performance of the classifier can be used as an indication of which region carries more information relevant to the task. If in spatial tasks classifiers are more accurate when restricted to the hippocampus than when restricted to the perirhinial cortex, this would indicate that the former is the locus of spatial processing. The conclusion can also be more graded. For example, the hippocampus might carry detailed spatial information but also some impoverished item information, and the perirhinial cortex might carry detailed item information but also some impoverished spatial information.

fine-grained computational diversity of neurons and neural networks within a given brain region, such as the premotor cortex. As we saw, if allowed to build neural networks composed of units with that much computational diversity, you can come up with processes that look like the simulations of ST or like the categorization and interface processes of TT. So coarse or even fine-grained activation in these areas in **execution** and **observation** cannot, by itself, be used to determine which of these processes is actually implemented. As illustrated by the recognition memory example, MVPA is extremely well suited for these sorts of cases. Once you train a classifier to learn, e.g., the sorts of multivoxel patterns associated with paradigmatic simulation processes, and the ones associated with paradigmatic abstract categorization processes, it can then tell you whether in some particular condition (e.g., **observation**), these diverse computational micro-units are working together in a way that is more like a simulation process or more like a categorization process.

4 Conclusion

There's been something of a theoretical, scientific, and popular backlash against the promises and aspirations of Cognitive Neuroscience.[26] This reaction is not unjustified. In the case of Philosophy, practices and opinions vary wildly. For every paper that claims that emotion-circuit activation in moral tasks undermines deontological ethics, or that sensory-motor activation in language tasks undermines classical cognitivism, you can find another one that proceed as if neuroscience has not taught us anything about 'higher' capacities such as decision making and language processing. This discussion of reverse inference supports a measured optimism. Careful analysis of the case of mirror neurons and ST seems to support, yet again, the position of the neuro-skeptics. ST theorists make a common mistake: they fail to meet the linking condition of reverse inference. Respecting this condition is extremely hard for any L-reverse inference. However, the linking condition can be directly satisfied by studies based on MVPA. This technique raises a host of new challenges; but these are not the old suspects of 'lack of selectivity', 'excessive functional locationism', and other charges best summed by labeling Cognitive Neuroscience as the 'New Phrenology'. As L-reverse inference becomes a tool of the past, and as these familiar objections loose much of their relevance, it is crucial that Philosophers of Mind and Science turn their attention to the new and much more powerful pattern decoding techniques.

[26] In philosophy, Jerry Fodor has been consistently critical of the alleged impact of neuroscience on the study of higher cognition (Fodor, 1974, 1997, 1999). Some of his concerns have been carefully developed by Coltheart (2006a,b); Trssoldi et al. (2012); Coltheart (2013). Several neuroscientists have expressed concern about the ways in which neuroscientific data has been used to make grand claims about cognition (see, e.g., some of the discussions reported in Miller (2008)). The more popular concerns are captured in various popular blogs and recent books such as Satel and Lilienfeld (2013).

References

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

Caramazza, A., S. Anzellotti, L. Strnad, and A. Lingnau (2014). Embodied cognition and mirror neurons: A critical assesment. *Annual Review of Neuroscience 37*, 1–15.

Catmur, C., V. Walsh, and C. Heyes (2007). Sensormotor learning configures the human mirror neuron system. *Current Biology 17*(17), 1527–1531.

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), 233–252.

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 child's theory of mind. *Cognition 44*, 283–296.

Fodor, J. A. (1997). Special sciences: Still autonomous after all these years. *Nous 31*, 149–63.

Fodor, J. A. (1999). Let your brain alone. *London Review of Books 21*.

Fogassi, L., P. F. Ferrari, S. Gesierich, B. an Rozzi, F. Chersi, and G. Rizzolatti (2005). Parietal lobe: From action organization to intention understanding. *Science 308*, 662–667.

Gallese, V. and A. Goldman (1998a). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences 2*(12), 493–501.

Gallese, V. and A. Goldman (1998b). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences 2*(12), 493–501.

Goldman, A. (2009). Mirroring, mindreading, and simulation. In J. A. Pineda (Ed.), *Mirror neuron systems*, pp. 311–330. Humana Press.

Goldman, A. I. (2012). A moderate approach to embodied cognitive science. *Review of Philosophy and Psychology 3*, 71–88.

Gopnik, A. and H. M. Wellman (1992). Why the child's theory of mind really is a theory. *Mind & Language 7*(1-2), 145–171.

Greene, J. (2009). The cognitive neuroscience of moral judgment. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences* (4th ed.)., Chapter 68, pp. 987–999. Cambridge, MA: MIT Press.

Greene, J., R. Sommerville, L. Nystrom, J. Darley, and J. Cohen (2001). An fMRI investigation of emotional engagement in moral judgment. *Science 293*, 2105–08.

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: W. W. Norton and Company.

Hutzler, F. (2013). Reverse inference is not a fallacy per se: Cognitive processes can be inferred from functional imaging data. *Neuroimage in press*.

Iacobani, 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. (2008). *Mirroring people: The new science of how we connect with others*. New York: Farrar, Straus and Giroux.

Iacoboni, M., I. Molnar-Szakacs, V. Gallese, G. Buccino, J. C. Mazziotta, and G. Rizzolatti (2005). Grasping the intentions of others with one's 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.

Knutson, B., E. G. Wimmer, S. Rick, N. G. Hollon, D. Prelec, and G. Loewenstein (2008). Neural antecedents and the endowment effect. *Neuron 58*, 814–22.

Lindquist, K. A., T. D. Wager, K. H., B.-M. E., and B. L. F. (2012). The brain basis of emotion: a meta-analytic review. *Behavioral and Brain Sciences 35*, 121–202.

Machery, E. (2013). In defense of reverse inference. *British Journal for the Philosophy of Science* (published online).

Miller, E. K. and J. D. Cohen (2001). An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci. 24*, 167–202.

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

Mole, C. and C. Klein (2010). Confirmation, refutation, and the evidence of fmri. In S. J. Hanson and M. Bunzl (Eds.), *Foundational Issues in Human Brain Mapping*, Chapter 9, pp. 99–112. Cambridge, MA: The MIT Presss.

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.

Phelps, E. (2006). Emotion and cognition: insights from studies of the human amygdala. *Annual Review of Psychology 57*, 27–53.

Phelps, E. (2009). The study of emotion in neuroeconomics. In P. W. Glimcher, C. F. Camerer, E. Fehr, and R. A. Poldrack (Eds.), *Neuroeconomics: Decision Making and the Brain*, Chapter 16, pp. 233–250. London: Academic Press.

Phelps, E. and M. Delgado (2009). Emotion and decision making. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences*, Chapter 76, pp. 1093–1105. Cambridge, MA: MIT Press.

Pietroski, P. M. (2010). Concepts, meanings, and truth: first nature, second nature, and hard work. *Mind & Language 3*, 247–278.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences 10*(2), 59–63.

Poldrack, R. A. (2008). The role of fmri is 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.

Ramachandran, V. S. (2000). Mirror neurons and imitatoin learning as the driving force behind the "great leap forward" in human evolution.

Rizzolatti, G. and L. Craighero (2004). The mirror-neuron system. *Annual Review of Neuroscience 27*, 169–192.

Rizzolatti, G., L. Fadiga, V. Gallese, and L. Fogassi (1996). Premotor cortext and the recognition of motor actions. *Cognitive Brain Research 3*, 131–141.

Rizzolatti, G., L. Fogassi, and V. Gallese (2001). Neurophysiological mechanisms underlying the understaning 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. 625–640. 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). Deconing patterns of human brain activity. *Annual Review of Psychology 63*, 483–509.

Trssoldi, 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*(155-165).

Yarkoni, T., R. A. Poldrack, T. E. Nichols, D. C. Van Essen, and T. D. Wager (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods 8*(8), 665–70.