Mapping the Mind: Bridge Laws and the Psycho-Neural Interface

Marco J. Nathan and Guillermo Del Pinal*

Synthese, Vol. 193, pp. 637-57, 2016
The final publication is available at Springer via
http://dx.doi.org/10.1007/s11229-015-0769-2

Abstract

Recent advancements in the brain sciences have enabled researchers to determine, with increasing accuracy, patterns and locations of neural activation associated with various psychological functions. These techniques have revived a longstanding debate regarding the relation between the mind and the brain: while many authors claim that neuroscientific data can be employed to advance theories of higher cognition, others defend the so-called 'autonomy' of psychology. Settling this significant issue requires understanding the nature of the *bridge laws* used at the psycho-neural interface. While these laws have been the topic of extensive discussion, such debates have mostly focused on a particular type of link: *reductive laws*. Reductive laws are problematic: they face notorious philosophical objections and they are too scarce to substantiate current research at the intersection of psychology and neuroscience. The aim of this article is to provide a systematic analysis of a different kind of bridge laws—*associative laws*—which play a central, albeit overlooked role in scientific practice.

1 Introduction

In a classic paper attempting to undermine theoretical reductionism, Jerry Fodor (1974, p. 97) noted that "the development of science has witnessed the proliferation of specialized disciplines at least as often as it has witnessed their reduction to physics, so the widespread enthusiasm for reduction can hardly

*Both authors contributed equally to this work. We are grateful to Max Coltheart, Kateri McRae, Bruce Pennington, and three anonymous reviewers for constructive comments on various versions of this essay. Some of the ideas developed here were presented at the Neuroscience Research Group at the University of Denver, at the 2014 Annual Conference in History and Philosophy of Science at the University of Colorado at Boulder, and at the 2014 Meeting of the Philosophy of Science Association in Chicago: all audiences provided helpful feedback.

be a mere induction over its past successes." Four decades later, Fodor's assessment remains accurate; indeed, it has been reinforced. Rather than being progressively reduced to physics, the special sciences have sprawled into a number of burgeoning subfields. Yet, at the same time, we have also witnessed the rise of *interdisciplinary* studies. If, as Fodor holds, the special sciences are relatively 'autonomous,' what explains the recent proliferation of fields such as neurolinguistics, moral psychology, and neuroeconomics?

The relation between different scientific fields has been extensively debated in philosophy and the particular case of psychology and neuroscience has gathered enormous attention. As reported in Bourget and Chalmers (2013), the dominant position is now *non-reductive physicalism*—the thesis that, although mental states are realized by brain states, mental kinds cannot, in general, be reduced to neural kinds. As we discuss below, this position fails to adequately address an important issue, namely, why studying the brain can inform our understanding of the mind. The failure to provide a convincing answer to this question is especially troublesome given the current trend in cognitive neuroscience, where advancements in neuroimaging have begun to affect theories of higher cognition, such as language processing and decision making (Gazzaniga 2009; Mather et al. 2013; Glimcher and Fehr 2014). If theorists are right that the mapping of mental kinds onto neural kinds is too problematic to substantiate any meaningful interaction at this interface, is neuroscience simply promising something that cannot be achieved? Or does the constant use of neural data in fields such as neurolinguistics and neuroeconomics mean that philosophical critique misunderstands the relation between cognitive and neural levels?

In this article, we argue that the tension between meta-theory and scientific practice stems from the failure to distinguish between different types of bridge laws, that is, principles that link kinds across domains. On the one hand, theorists have generally been concerned with *reductive* bridge laws. On the other hand, most bridge laws currently employed in cognitive neuroscience are not reductive; they are *associative* statements that are categorically distinct from the contingent type-identities typically employed in derivational reduction and in more recent reductive approaches. The aim of this essay is to provide an account of associative bridge laws which, despite their widespread use in neuropsychology, have never been systematically discussed. We begin by introducing the role of bridge laws in traditional models of derivational reduction and rehearsing some well-known problems (§2). Next, we present the kind of bridge laws that are employed in neuroscientific studies of higher-cognition (§3) and elucidate the main differences between these associative statements and their reductive counterparts (§4). We conclude by discussing some implications of our analysis for extant debates in the philosophy of mind and science (§5).

Before we begin, two brief remarks about terminology. First, we employ the term 'bridge law' as referring to any statement that maps predicates across theories or domains of science. Depending on the nature of the interfield relation, these links can assume different forms. Whereas reductionist accounts require reductive laws, different types of bridge laws can be found in non-reductive theories. Second, we shall not enter longstanding metaphysical debates on the

notions of *event* and *natural kind*. For present purposes, we treat natural kinds as predicates that fall under the laws or generalizations of a branch of science (Fodor 1974). Similarly, a P-event is an event involving property P.

2 Bridge Laws in Theory Reduction

In what became a *locus classicus*, Nagel (1961) characterized reduction as a deductive derivation of the laws of a reduced theory P from the laws of a reducing theory N. Such derivation requires that the predicates of P be expressed in terms of the predicates of N. For instance, suppose that we want to show that a law $L_P : P_1x \rightarrow P_2x$, expressed in the language of theory P, can be reduced to—that is, derived from—a law $L_N : N_1x \rightarrow N_2x$, expressed in the language of theory N. (For the sake of simplicity, let us assume that the languages of the two theories do not overlap, i.e., that the predicates of P do not also belong to N, and vice versa.) What we need is a series of *bridge laws*, that is, principles that govern the translation of the relevant P-predicates into N-predicates:

(R_1) $P_1x \leftrightarrow N_1x$

(R_2) $P_2x \leftrightarrow N_2x$

How should the '$\leftrightarrow$' connective be interpreted, in order for R_1 and R_2 to play their role in Nagelian reduction? Fodor (1974) makes a number of important points. First, '$\leftrightarrow$' must be transitive: if P_1 is reduced to N_1, and N_1 is reduced to Q_1, then P_1 is thereby reduced to Q_1. Second, '$\leftrightarrow$' cannot be read as 'causes,' for causal relations tend to be asymmetric—causes bring about their effects, but effects generally do not bring about their causes—whereas bridge laws are symmetric: if an P_1-event is a N_1-event, then a N_1-event is also an P_1-event. Given these two features, bridge laws are most naturally interpreted as expressing *contingent event identities*. Thus understood, R_1 can be read as stating that P_1 is *type-identical* to N_1.[1]

Note that the Nagelian model of reduction provides a clear-cut account of how discoveries at the neural level could, in principle, inform theories of higher cognition. Suppose that we want to test hypothesis $L_P : P_1x \rightarrow P_2x$, which posits a law-like connection between two psychological predicates P_1 and P_2. If we had a pair of reductive bridge laws that map P_1 and P_2 onto neural kinds N_1 and N_2, respectively, then we could confirm and explain the law-likeliness of L_P directly by uncovering the neural-level connection $L_N : N_1x \rightarrow N_2x$. This is because, as noted above, the bridge laws employed in derivational reduction express type-identities. Consequently, if P_1 and P_2 are type-identical to N_1 and N_2 and there is a law-like connection between P_1 and P_2, there will also be

[1] As Fodor notes, reductive bridge laws express a stronger position than *token physicalism*, the view that all events that fall under the laws of some special science are physical events. Statements such as R_1 and R_2 presuppose *type physicalism*, according to which every kind that figures in the laws of a science is type-identical to a physical kind. Since our focus is not on physicalism *per se*; the relevant claim is whether the kinds of one science can be reduced to the kinds of a more fundamental science, not necessarily to physics.

an analogous law-like connection between N_1 and N_2. To illustrate, consider the following analogy. Suppose, for the sake of the argument, that *water* and *sodium chloride* can be reduced, in the sense of being type-identical, to H_2O and $NaCl$. If one provided a successful explanation of why $NaCl$ dissolves in H_2O, under specific circumstances, then one has thereby explained why sodium chloride dissolves in water, under those same conditions. In short, the reductive model suggests a specific goal for cognitive neuropsychology, namely, to look for neural-level implementations of psychological processes, which can then be used directly to test and explain psychological laws.

The well-known objection against type-physicalism is that natural kinds seldom correspond so neatly across levels. Although one could make a case that heat is reducible to mean molecular kinetic energy, or action-potentials to nerve impulses, the reigning consensus in philosophy of science is that contingent event identities are too scarce to make derivational reduction a plausible general inter-theoretic model (Horst 2007). In most cases, there seem to be no physical, chemical, or macromolecular kinds that correspond to biological, psychological or economic kinds, in the manner required by the reductionist scheme. This, simply put, is the *multiple-realizability argument* against the classical model of derivational reduction (Putnam 1967; Fodor 1974). The basic idea is that instead of R_1 and R_2, what we usually find are linking laws such as R_3, which capture how higher-level kinds can be potentially realized by a variety of lower-level states:

(R_3) $P_1x \leftrightarrow N_1x \vee \ldots \vee N_nx$

In response to the multiple-realizability argument, philosophers pursued two alternative routes, depending on their metaphysical inclinations. One strategy consists in refining the reductive framework. This can be done in various ways, for instance, by relativizing Nagelian bridge-laws to types of physical systems or individuating psychological and neural kinds more finely (see §4.3), or by trying to avoid altogether any commitment to bridge laws (Hooker 1981; Bickle 1998; Kim 1999, 2005).[2] Following a different path, many philosophers embraced a functionalist approach, according to which mental states are individuated by their causal roles, independently of their physical realization (Putnam 1967; Fodor 1974, 1997). Psychofunctionalists embrace the multiple realizability of higher-level states: on the standard functionalist reading of R_3, psychological kind P_1 can be realized by a variety of neural kinds N_i. Hence, functionalism leaves open one way in which neuroscience can contribute to psychology: since, according to R_3, P_1 is token-identical to one of its neural realizers, the presence of *any* N_i would be evidence for the engagement of psychological kind P_1 in a specific task. However, this approach suggests that neuroscience can be only applied to psychology when the neural realizer(s) of cognitive states are known—an extremely demanding presupposition, given our current knowledge.

[2]However, it has been persuasively argued that any form of *bona fide* reductionism requires some kind of bridge laws (Marras 2002; Fazekas 2009).

Let us take stock. Derivational reduction provides a clear explanation of how neuroscience can be used to advance psychological theories, but it presupposes an implausible and overly-demanding account of the linkage of kinds across domains. Functionalism, in contrast, avoids the unpalatable assumptions of reductionism and suggests a subtler way in which neuroscientific evidence can contribute to psychological debates. Yet, the standard functionalist model is still extremely demanding, as it requires bridge laws mapping psychological states onto some their neural realizers, in the manner illustrated by R_3.

Part of the problem with the extant debate, we surmise, is that reductionists and functionalists alike share an overly restrictive view of the psycho-neural interface. Researchers in both camps often talk as if the only potential contributions of neuroscience to psychology are:

(i) To establish *correlations* between cognitive- and neural-level events, e.g., to find the brain locations *where* particular mental functions are computed.

(ii) To discover the neural-level mechanisms that *compute/implement* cognitive processes, i.e., to establish *how* the brain actually computes/implements specific mental functions.

That neuroscience can contribute to project (i) is hardly controversial; the problem is that, by itself, (i) seems pointless, since seeking mind-brain correlations that do not contribute to an explanation of *how* neural mechanisms compute cognitive functions becomes a sterile vindication of token physicalism. Therefore, it is common to assume that (i) is valuable only insofar as it contributes to the more substantial and ambitious project (ii). Is neuroscience currently at the point of uncovering the mechanisms that implement and compute mental functions in the brain? Reductionists tend to stress the remarkable successes in discovering neural mechanisms of sensory systems, such as early vision, pain, taste, and other basic sensations (Bickle 2003; Kim 2006). Antireductionists, in contrast, emphasize that comparable achievements cannot be claimed for language processing, decision making, and other functions of higher cognition and, consequently, deem the pursuit of project (ii) hopeless (Fodor 1999) or, at best, drastically premature when applied to the more central cognitive systems (Gallistel 2009; Coltheart 2013).

This view of the psycho-neural interface, assumed by reductionists and functionalists alike, is too restrictive. In the rest of this article, we argue that neuroscientific data can be fruitfully employed to advance psychological theories, even in the absence of strongly reductive bridge laws such as R_1 and R_2, which type-identify kinds across levels, or weaker statements, such as R_3, expressing the multiply-realizable token-identities of psychological kinds at the neural level. In order to capture the success of these interdisciplinary studies, we need a novel account of bridge laws that captures their non-reductive character and explains how they can be applied even when the neural realizers are unknown. To flesh-out the nature of these links, we focus on one of the main techniques which cognitive neuroscientists use to make neural data and theories bear on cognitive-level hypotheses: *reverse inference*.

3 Bridge Laws and Reverse Inferences

In order to discriminate between competing cognitive hypotheses, neuroscientists often 'reverse infer' the engagement of a cognitive state or process, in a given task, from particular locations or patterns of brain activation (Henson 2005; Poldrack 2006; Del Pinal and Nathan 2013; Hutzler 2014; Machery 2013). These *reverse inferences* presuppose the availability of bridge laws; yet, contrary to a widespread assumption, the required links are not reductive, they are what we call *associative bridge laws*. In this section, we examine the role of bridge laws in two kinds of inferences employed in neuroimaging studies: *location-based* and *pattern-based reverse inferences*. More specifically, we focus on studies of decision-making—a paradigmatic domain of higher-cognition—aimed at discriminating between the processes which underlie behavioral generalizations.

To begin, consider the following psychological generalization, somewhat simplified for the sake of illustration, where s ranges over 'normal' adults:

(G_M) If s is faced with the option of performing an action a that will result in the death of fewer people than would die if s were not to perform a, s will choose a unless doing so requires using a person directly as a means.

G_M captures a distinctive capacity of higher-cognition which is in need of explanation. We shall refer to the level at which we isolate these types of psychological generalizations as *Marr-level 1*.[3] Given a Marr-level 1 generalization, one can explore the underlying cognitive processes: such conjectures are usually referred to as *Marr-level 2 hypotheses*. Consider two competing explanations of G_M:

(M) In moral decision making, subjects generally follow consequentialist rules. However, in cases which involve using another person directly as a means, consequentialist rules are overridden by *negative emotions*.

(M^*) In moral decision making, subjects generally follow consequentialist rules. However, in cases which involve using another person directly as a means, consequentialist rules are overridden by *deontological rules*.

M and M^* are very different explanations of G_M. Whereas M explains the behavioral pattern as a conflict between rules and emotions, M^* explains the same pattern as a conflict between different types of rules: consequentialist vs. deontological.

M and M^* are competing Marr-level 2 hypotheses about the cognitive processes which underlie a Marr-level 1 generalization. To adjudicate between them, researchers use reverse inferences, which require two preliminary steps. First, the competing processes must be functionally decomposed, for entire processes

[3] In an influential discussion, Marr (1982) argued that information-processing systems should be investigated at three complementary levels. Hypotheses at Marr-level 1 pose the computational problem: they state the task computed by the system. Hypotheses at Marr-level-2 state the algorithm used to compute Marr-level 1 functions: they specify the basic representations and operations of the system. Finally, hypotheses at Marr-level 3 specify how Marr-level 2 algorithms are implemented in the brain: they purport to explain *how* these basic representations and operations are realized at the neural level.

such as M and M^* are too coarse-grained to be directly mapped onto patterns or regions of neural activation. Next, the subcomponents of the competing processes for which there are bridge laws must be identified. To illustrate, let us assume that, in task T, cognitive process M posits the engagement of subprocess m_1, whereas M^* posits the engagement of subprocess m_1^*, and that $m_1 \neq m_1^*$. Further, suppose that we have the following bridge laws connecting m_1 and m_1^* with regions or patterns of neural activation n_1 and n_1^*:

(A_1) $m_1 \otimes n_1$

(A_2) $m_1^* \otimes n_1^*$

Note that '$\otimes$' is different from the '$\leftrightarrow$' connective figuring in reductive bridge laws. We shall discuss the basic properties of such relation in §4. The important point here is simply that '$\otimes$' stands for an associative relation that allows one to reliably infer the presence of one relatum from the other.

To illustrate the application of statements such as A_1 and A_2, consider some bridge laws used to discriminate between M and M^*. Assume that m_1 stands for processes involving negative emotions such as fear, and that m_1^* stands for ruled-based processes such as following simple instructions. Researchers have established a close connection between processes involving negative emotions and activation in certain neural regions such as the amygdala and the ventromedial prefrontal cortex (VMPFC).[4] This connection is captured by A_1. Researchers have also established a connection between rule-based and controlled reasoning and activation in the dorsolateral prefrontal cortex (DLPFC).[5] A_2 captures this connection by associating m_1^* with activation in the DLPFC.

Given A_1 and A_2, one can devise neuroimaging experiments to discriminate between M and M^*. For example, Greene and colleagues (2001) scanned subjects making moral decisions in two sets of tasks that involve choosing whether to sacrifice one innocent person to save five, as in the famous trolley problems. The relevant difference is that in one set of tasks all the choices that would save five people involve using another person directly as a means (*personal cases*), whereas in the other set subjects can save five by sacrificing one indirectly, that is, without using the person as a means (*impersonal cases*).[6] Greene and

[4] In general, the amygdala is critically involved in conditioned and unconditioned fear response in animals, including humans. For example, patients with selective damage to the amygdala show no physiological response to a previously fear-conditioned stimulus, although they can explicitly remember the conditioning experience (Kandel et al. 2013, Ch. 48).

[5] Miller and Cohen (2001) present several studies that support the key role of the DLPFC in cognitive control and rule-guided processes. A relevant set of experiments are based on the famous Stroop task, in which subjects are instructed to name the color of the ink of words as they appear on a screen. Famously, reaction times and error rates increase dramatically when subjects read color-terms that differ from the color of their ink. Miller and Cohen present imaging studies which show that, in the misleading cases, subjects who manage to follow the correct rule and name the word's ink color showed increased activation in DLPFC, compared to subjects who fail the task.

[6] In the classic version of the trolley problem, personal cases are exemplified by the 'footbridge' scenario, where five people are saved by throwing a corpulent person on the track. Impersonal cases are exemplified by the 'switch' scenario, where five people are saved by pulling a lever that diverts the trolley onto a parallel track where it will kill a single person.

colleagues found that, relative to impersonal cases—and to structurally analogous non-moral control tasks—personal cases result in differential activation of the amygdala and VMPFC, and less activation of DLPFC. Given that A_1 associates amygdala activation with negative emotions, and that A_2 associates DLPFC activation with rule-based and controlled reasoning, this finding favors M over M^*. This is because, according to M, in personal cases, decisions not to sacrifice one person to save five are based on negative emotions. In addition, M predicts that areas involved in rule-based reasoning should be more active in impersonal compared to personal cases. In contrast, M^* incorrectly predicts that personal and impersonal cases should engage rule-based areas equally, since both cases involve applying different types of rules.

Critics of the relevance of neuroimaging experiments for psychology often assume—more or less explicitly—that all bridge laws currently employed in reverse inferences associate cognitive processes with *locations* of neural activation, as in the previous example. However, this is a mistake: in some cases, the relevant bridge laws map cognitive processes onto particular *patterns* of neural activation. Indeed, pattern-based inferences, which are rapidly becoming one of the main ways of studying cognition, have significant implications for the psycho-neural interface. To see this, let us examine an example from the study of recognition memory. Consider the following generalization:

(G_N) 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.

Among the Marr-level 2 explanations of G_N recently advanced in episodic memory research, two of the most important are the following:

(N) Recognition decisions are based on two processes which draw on 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.

(N^*) Recognition decisions are based on two processes which draw on 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*.

While N and N^* agree on the basic components underlying recognition decisions, they posit different interactions. According to N, subjects generally use recollection information to decide whether items are old, and only rely on intuitions of familiarity when such information is unavailable. In contrast, N^* predicts that certain contextual cues will induce subjects to make familiarity-based recognition decisions even if recollection information is available.

In pattern-based recognition studies, 'classifiers' are trained to determine the multi-voxel patterns associated with recollection and familiarity processes. Specifically, classifiers are trained in tasks where experimenters can control

which cognitive process is engaged. For instance, in an experiment designed to adjudicate between N and N^*, 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 in 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 makes this method especially interesting is that the reliability of the classifiers can be established within the experiment itself. 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, where experimenters control which process is engaged, establishes the bridge laws that will then be used in reverse inferences.

Having obtained the relevant bridge laws which map recollection and familiarity onto multi-voxel patterns, one can then test competing hypotheses N and N^* regarding the dynamics underlying recognition-decisions in cases where the engagement of the sub-processes cannot be directly controlled. For example, 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 ('ball'), unrelated lures ('horse'), and previously unstudied switch-plurality lures ('balls'). 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. Hence, according to hypothesis N, the classifier should categorize the corresponding voxel patterns as recollection processes (since this is the default). In contrast, N^* predicts that the classification should be more variable, involving—at least in some cases—familiarity processes. Experimental results support N^* over N: when both types of information are available, various contextual cues determine whether subjects use familiarity or recollection as the basis of their recognition decision (Norman et al. 2009).

4 Associative Bridge Laws

The previous examination of reverse inference allowed us to place associative bridge laws such as A_1 and A_2 in their context of use. The aim of this section is to make explicit the characteristic features of these linking statements. As we shall see, unlike their reductive counterparts, associative bridge laws are *probabilistic* and *context-sensitive* relations that do *not identify* their relata, either at the type-level or at the token-level.

4.1 Probabilities

The first main feature of associative bridge laws is their *probabilistic* nature. To clarify, consider a recent debate about the 'selectivity' of brain regions. Several critics have emphasized that the success of a reverse inference depends on the degree of selectivity of the relevant brain regions (Uttal 2002; Ross 2008; Phelps 2009; Anderson 2010; Coltheart 2013). Suppose that some bridge law maps neural activation in n_1 onto the engagement of cognitive process m_1. According to critics, this linkage allows one to legitimately reverse infer the engagement of m_1 from the activation of n_1 only provided that region n_1 activates for the cognitive process of interest, in this case n_1, *and no other*. This is because, the objection runs, if n_1 also activates when m_2, m_3, and m_4 are engaged, one *cannot* reverse infer to m_1 merely on the evidence of n_1 activation. The problem is that there is widespread consensus among cognitive neuroscientists that very few brain regions are *maximally selective* in the sense just described. From this perspective, then, it looks like most reverse inferences are actually invalid, as they rely on an unjustified assumption of maximal selectivity.

This substantial and influential worry ought to be addressed with care. For starters, it is undeniable that few, if any, brain regions are indeed maximally selective. Hence, virtually no brain region can be mapped onto cognitive functions via a *single* bridge law; rather, each brain region is covered by *multiple* bridge laws which associate it with a variety of cognitive functions. As a result, when we reverse infer the engagement of a cognitive function from the activation of a neural region, the inference falls short of absolute certainty. Confidence that one has identified the correct bridge law is a matter of degree, which is determined by the conditional probability that cognitive process m_1 is engaged, given activation in n_1. This conditional probability can be determined through a straightforward application of Bayes' theorem:

$$P(m_1|n_1) = \frac{P(n_1|m_1)P(m_1)}{P(n_1|m_1)P(m_1) + P(n_1|\neg m_1)P(\neg m_1)} \tag{1}$$

Equation (1) entails that the degree of belief in a reverse inference depends not only on the prior $P(m_1)$ but also on the selectivity of the neural response—i.e., on the ratio of the process-specific activation $P(n_1|m_1)$, to the overall likelihood of activation in that area across all tasks which do not involve m_1: $P(n_1|\neg m_1)$.

As an illustration, consider the example of moral decision making (G_M). As neuroscientists note, the amygdala is also activated by processes that are not related to negative emotions in any obvious way. Consequently, amygdala activation does not deductively entail the engagement of fear or related negative emotions. However, it does not follow that inferences from amygdala activation to the presence of negative emotions are invalid; what follows is simply that such inferences are *inductive* or *probabilistic*. The case of the amygdala is not the exception, it is the norm: as noted, most brain regions are associated with multiple cognitive processes. Furthermore, this point is not restricted to location-based inferences, but also applies to pattern-based ones. The multi-

voxel patterns are, at best, a reliable guide for inferring (*via* bridge laws) the engagement of the associated cognitive process.

With all of this in mind, we can now turn to a refinement of the above critique, directed to the probabilistic nature of reverse inference. Several authors have argued that, since the application of a given bridge law in some task is determined by a conditional probability, most interesting reverse inferences turn out to be unacceptably weak (Miller 2008; Phelps 2009; Legrenzi and Umilta 2011). This objection underlies many skeptical claims about the use of reverse inferences and has led to the explicit suggestion that genuine progress at the psycho-neural interface requires reductionist bridge laws (Ross 2008; Anderson 2010). No doubt, in some cases, such accusations are justified: some proposed reverse inferences are indeed questionable, to say the least. Yet, this observation falls short of a general critique, for the significance of the lack of (maximal) selectivity on the validity of reverse inference has been substantially exaggerated. This is because critics often overlook another important characteristic of associative bridge laws, namely, their *context sensitivity*.

4.2 Context-Sensitivity

In an influential article, Poldrack (2006) noted that the conditional probability that a cognitive state or process m_1 is engaged given a neural state or process n_1 should be determined *relative to a particular task*. However, to simplify the discussion, Poldrack intentionally ignored this task-relativity in the rest of his analysis. That simplification had the unfortunate consequence that several ensuing discussions also ignored the task-relativity of bridge laws in reverse inferences. This resulted in a misleading objection.

Consider, again, the selectivity of the amygdala, which plays a central role in several studies of decision making. Although the amygdala is typically involved in processing fear and other negative emotions, it is also involved in many other cognitive processes that are usually unmentioned in studies such as Greene et al. (2001). Such processes include the perception of odor intensity, sexually arousing stimuli, and trust from faces (Phelps 2006; Lindquist et al. 2012), as well as the processing of faces from other races, and the perception of biological motion and sharp contours (Phelps 2009). It has also been claimed that the main function of the amygdala is to process novel or emotionally salient stimuli—not fear-related stimuli *per se* (Lindquist et al. 2012). Based on these considerations, Phelps (2009) argues that amygdala activation in a given psychological task could signal the engagement of *any* of these cognitive processes. Consequently, reverse inferences such as the ones used by Greene and colleagues overestimate the conditional probability that negative emotions are engaged, given amygdala activation.

What Phelps and other critics (e.g., Klein 2011) overlook is that the probability that a particular bridge law applies, given the activation of a brain region, should be determined relative to relevant tasks. Specifically, in the case under consideration, the success of the reverse inference does *not* depend on the assumption that we can reliably infer the engagement of negative emotions from

differential activation in the amygdala. What is required is that the engagement of negative emotions can be inferred from the pattern of neural activation observed *in the particular task under consideration* (Del Pinal and Nathan 2013). In other words, the inference is from differential amygdala-activation *in personal scenarios* to the engagement of negative emotions. Once the inference is framed in these terms, we can see that most other cognitive processes that also involve the amygdala are not plausible explanations for such differential activation, and can thus be ruled out. Consider, for instance, the tasks used by Greene and colleagues (2001). Personal cases do not differ from impersonal ones with respect to stimuli related to odor, facial-processing, sexual material, sharp-contours, or the comparative novelty of the tasks. Hence, relative to personal cases, the conditional probability of the engagement of negative emotions, given amygdala activation, is significantly higher than suggested by the objection presented above.[7]

How, precisely, to formalize the task-relativity of reverse inferences is a subject of current debate. A promising proposal is to incorporate task-relativity into the Bayesian formula (Hutzler 2014). The main idea is to revise (1) by conditionalizing explicitly on the relevant task—call it 't_1.'

$$P(m_1|n_1 \wedge t_1) = \frac{P(n_1|m_1 \wedge t_1)P(m_1|t_1)}{P(n_1|m_1 \wedge t_1)P(m_1|t_1) + P(n_1|\neg m_1 \wedge t_1)P(\neg m_1|t_1)} \quad (2)$$

To motivate (2), Hutzler presents a simple thought experiment. Imagine that activation in the left fusiform gyrus (n_1) is covered by two bridge-laws: A_1 associates n_1 with access to the mental lexicon and A_2 associates n_1 with face perception. Assume that a visual word-presentation-task t_1 results in n_1. The question is whether n_1 significantly increases one's confidence that t_1 involves access to the mental lexicon. Equation (1) entails that confidence in t_1 involving access to the mental lexicon is decreased by A_2, according to which n_1 could also signal face perception. But this is counterintuitive, for t_1 clearly has nothing to do with face perception. In contrast, Equation (2) makes A_2 irrelevant: by taking t_1 into account, we can eliminate the possibility that *in this case* n_1 underlies face perception. In short, conditionalizing on tasks increases the strength of actual location-based inferences: decreasing the number of possible cognitive functions that the neural measure could be signaling reduces the risk of false alarms. This is essentially the same strategy we followed in defending the reverse inference from amygdala activation to engagement of emotions in the moral decision making experiments.[8]

[7] We surmise that the task relativity of reverse inferences is systematically overlooked because methodological discussions (e.g., Poldrack 2006; Phelps 2006) often consider only arbitrary 'empty' tasks which do not eliminate any processing possibilities (that is, any bridge laws) for the brain region of interest. Hence, reverse inferences seem intuitively weak. However, once we consider the tasks relevant to each reverse inference, we can eliminate some subset of bridge laws which cover the brain regions of interest, thereby increasing their strength.

[8] An alternative is to reformulate reverse inference in likelihoodist terms (Machery 2013). Consider two competing cognitive hypotheses m_1 and m_2 and neural activation data n_1. On this view, n_1 favors m_1 over m_2 if and only if $P(n_1|m_1) > P(n_1|m_2)$. In a likelihoodist

Of course, computing Equation 2 is not always a straightforward matter. Consider again the moral decision making case. While sexually arousing stimuli and sharp contours can be ruled out right away, the possibility that differential amygdala activation might result from the increased presence of biological motion cannot be so easily dismissed. In such cases, a follow-up task should be devised to test for the remaining possibilities. In general, there is no formal or purely objective way to determine which cognitive subprocesses are (ir)relevant, given a certain task, and the role of 'value judgments' in scientific practice is well known (Kuhn 1977). Sometimes the decision is obvious; when it is not, further experimental work becomes necessary.

It is crucial to note that the criticisms of reverse inference based on lack of selectivity are typically raised against location-based inferences. In contrast, pattern-based reverse inferences provide an elegant solution to this problem. This is one reason why multivariate pattern-decoding methods are now generally regarded as superior techniques to univariate location-based methods. The key advantage of decoding techniques is that the reliability of classifiers can be established within the experiment itself. In the recognition example from §3, this was done by saving a subset of the recollection and familiarity blocks for later testing (so these data sets are not part of the classifier training sets). In this phase, experimenters can control whether familiarity or recollection processes are engaged in the task. One can then feed these neural data sets to the classifier, which maps them onto recollection or familiarity processes. Such predictions are compared to the original labels to determine their accuracy. For theorists interested in reverse inference, the classifier will only be useful if its accuracy is significantly above chance.[9] This condition was satisfied in the recognition example. In other cognitive domains—e.g., visual perception, phonological processing, and decision making—classifier accuracy can be extremely high (for an overview, see Tong and Pratte 2012). In short, when we use pattern-decoding techniques we can quantitatively estimate the degree to which a classifier can use patterns of brain activation to predict the engagement of specific mental processes in some task. This results in a more formal

framework, one only compares cognitive hypotheses that are under dispute, treating reverse inference as an inherently comparative technique that tells us which among the competing hypotheses is favored by some neural evidence. One drawback of this suggestion is that evidence becomes purely comparative. On the other hand, the advantage of this approach is that the relevant likelihoods can be calculated without having to determine the base rates of activation of the brain regions involved. We should note that the main reason why Machery prefers this likelihoodist approach to reverse inference over the Bayesian account is that Equation 1 cannot, in general, be computed. This is because neuroscientists rarely know the base rates of activation of particular brain regions of interest. It is not clear whether Machery thinks this is still a serious problem for Huzler's refined Equation 2, since he does not directly consider that option. However, we should also note that there is a substantial literature on Bayesianism and imprecise probabilities that can be used to address Machery's concern (Joyce 2011).

[9] Cases where classifiers cannot perform significantly above chance can still be interesting, albeit for different reasons. Suppose that, in task t, a classifier underperforms when using data sets taken from some region n_1, but performs significantly above chance when using data from region n_2. This provides evidence that n_2 carries information relevant to performing t, whereas n_1 does not.

implementation of particular reverse inferences (Poldrack 2011).[10]

4.3 Non-Identity

Unlike their reductive counterparts, associative bridge laws do not presuppose any kind of identity—*a priori*, *a posteriori*, *necessary*, or *contingent*. To wit, in the moral decision making case, the bridge law mapping amygdala activation to the engagement of negative emotions presupposes neither the type-identity nor the token-identity of these two events. As we saw, the amygdala is differentially activated by a variety of cognitive processes that have little or nothing to do with negative emotions, and it might turn out that some unambiguously fear-or-distress-related processes are not accompanied by increased amygdala activation. We should make it clear that we are not recommending any departure from token-physicalism. Our point is simply that associative bridge laws are so metaphysically uncommitted that they would also be consistent with some rejections of token-physicalism.

A similar point applies to pattern-based inferences. Bridge laws used in the recognition case do not presuppose that recollection or familiarity processes are (type- or token-) identical to their associated multi-voxel patterns. For one thing, the patterns are only highly reliable—but not infallible—indicators of the corresponding processes. More importantly, even if we had perfect correlations, multi-voxel patterns are not plausible candidates for such identities. Voxel patterns are representations that average over the activation of thousands of neurons, but do not specify the actual neural mechanisms that compute cognitive-level processes. This, of course, is not to say that the possibility of a type-identity can be ruled out *a priori*: one might believe that, eventually, the neural mechanisms that carry out, say, recollection processes will be identified. However, this potential reduction is neither required nor presupposed by the use of pattern-based inferences to discriminate among competing hypotheses of the processes underlying recognition tasks.

To appreciate further what is distinctive about associative bridge laws, it is useful to contrast them with reductive accounts that respond to multiple realiz-

[10] Of course, this does not mean that there are no difficulties in using this method. Appropriate experimental design is crucial, especially since pattern classifiers are designed to use whatever information is available to make better predictions. In addition, there is still a question 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 stimulus and task (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).

ability by adding parameters or weakening Nagelian bridge laws. David Lewis (1969) famously argued that reductive type-identities are not meant to hold across the board. On his view, the bridge laws reducing mental states to brain states are implicitly restricted to a specific domain. For example, while pain *tout court* cannot be reduced to a single brain state, human pain, octopus pain, martian pain, etc. can each be reduced to a different type of brain state. Lewis' argument has been developed and refined by various philosophers (Hooker 1981; Enç 1983; Churchland 1986; Kim 1992) all of whom emphasize that contingent event identities should have as conditional antecedents some kind of domain restriction. To cite a textbook example, the standard identification of temperature with mean molecular kinetic energy in classical equilibrium thermodynamics is left completely unscathed, the arguments runs, by the observation that temperature is differently realized in gases, solids, vacuums, and other mediums. Relativizing or conditionalizing reductive bridge laws might ultimately lead to a substantial increase in their number. This is a topic of current debate. But what is important to note here is that associative bridge laws do not require restricted conditional *identities* of any kind. This is especially evident in the case of pattern-based inferences: the particular voxel patterns used to infer the engagement of each sub-type of recognition process—that is, the bridge laws—are often not even stable across individuals, let alone all human beings, and can only be used reliably in specific experimental conditions. In the recognition experiments, the voxel patterns were used by classifiers to infer the engagement of familiarity or recollection in a task where these processes were the only unknown variables. If a third task (say, a face-recognition process) were added, the pattern-classifier would have to be re-trained. In this case, there would be no guarantee that the patterns that were previously associated with familiarity and recollection, even if present, could still be used, in the new experimental settings, to reliably predict the original processes.

For similar reasons, associative bridge laws should also be distinguished from recent attempts to weaken Nagelian bridge laws by replacing type-identity with a condition of *connectability* based on co-referentiality. Klein (2009) argues that a higher-level science S is N-connectable to a lower-level science S' if and only if S' has the resources to introduce new terms, in its own vocabulary, which are co-referential with the predicates of S that are absent in S'. Choosing an account of reference determination in general, and then arguing for a particular case of co-referentiality, is a substantial endeavor that we can set aside. Associative bridge laws do not require that terms such as 'fear' and 'increased amygdala activation', or 'recollection' and 'voxel vector pattern V' be co-referential. All that matters is that the presence of the referent of one can be reliably inferred from the presence of the referent of another. The co-referentiality of terms in the relata of a bridge-law is consistent with—but not a necessary condition for—their successful employment in reverse inferences.

In short, the bridge laws which figure in location- and pattern-based reverse inferences do not assume any kind of identity between neural and cognitive states or processes. In order to play a role at the psycho-neural interface, associative bridge laws only need to allow us to reliably infer, in certain experimentally con-

trolled settings, the engagement of a cognitive state or process from particular locations or patterns of neural activation.

5 Implications

In the previous section, we analyzed the characteristic features of associative bridge laws by drawing on the way they are employed in scientific practice and contrasting them with their reductive counterparts. We now turn to their implications for various ongoing debates about inter-level relations in philosophy of mind and science. Specifically, we begin by discussing functional locationism and multiple realizability. We conclude by revisiting the traditional intrerpretation of Marr-levels and its relation to the alleged 'automony' of psychology.

5.1 Avoiding Radical Locationism

Many prominent scientists and philosophers argue that cognitive neuroscientists assume an unreasonably strong version of *functional locationism* (Van Orden and Paap 1997; Fodor 1999; Uttal 2001; Coltheart 2013; Satel and Lilienfeld 2013). Some have gone as far as labeling current cognitive neuroscience the 'new phrenology' (Uttal 2002). This critique often presupposes a reductive model of the psycho-neural interface. To wit, if one assumes both that bridge laws are reductive and that most reverse inferences are grounded in location-based neural data, then it becomes reasonable to conclude that cognitive neuropsychology is in the business of type-identifying cognitive functions with neural locations, blatantly ignoring multiple realizability and the failures of derivational reduction. While the charge of excessive functional locationism is sometimes warranted, it does not apply to properly conducted reverse inferences (Del Pinal and Nathan 2013). Furthermore, it ignores the current trend in cognitive neuroscience, at least if the increasing importance of pattern-based inferences is a reliable indicator (Poldrack 2008, 2011).

As illustrated by our examples, most reverse inferences do not associate the engagement of entire cognitive processes with specific locations of neural activation. The general strategy is to decompose the competing processes into their subcomponents and to consider those subcomponents that can be mapped, *via* bridge laws, to neural locations or patterns. We can then reliably reverse-infer the engagement of one of the cognitive processes, relative to a specific task. In the moral decision-making case, only one of the competing processes predicted the engagement of negative emotions in personal tasks, which is why differential amygdala-activation provided evidence in favor of M over M^*. The point to stress is that, for the argument to go through, one need not assume the functional localization of the entire moral decision-making processes. Pattern-based inferences are even less plausible targets for the charge of unjustified functional locationism. Classifiers use multi-voxel patterns to infer the engagement of recollection or familiarity in recognition tasks. Classifiers need not be given any location-related information, which allows, in principle, for the set of patterns

assigned to, say, recollection, to be implemented in various neural locations. Interestingly, recent studies suggest that key components of recognition processes are, indeed, functionally localized (Norman et al. 2010). To be sure, there remain several controversial issues regarding the foundations of cognitive neuropsychology, including the substantial question of how to formalize the context-relativity of reverse inference (Del Pinal and Nathan 2013; Hutzler 2014; Machery 2013). Yet, the wholesale dismissal of the entire cognitive neuropsychology of higher cognition as a 'sophisticated new phrenology in disguise' does not withstand serious scrutiny.

5.2 Accommodating Multiple Realizability

As discussed in §2, anti-reductionist philosophers maintain that the natural kinds of a 'higher' science cannot, in general, be reduced to kinds of a 'lower' science because natural kinds seldom correspond across domains in the way required by reductive bridge laws. A complete assessment of multiple realizability and reductionism lies beyond the scope of this article. Our point is simply that multiple-realizability, coupled with a reductive conception of bridge laws, generates serious problems for understanding the fruitfulness of the interdisciplinary work pursued in current neuroscience.

Associative bridge laws are perfectly consistent with the multiple-realizability of psychological kinds. Amygdala activation signals the engagement of processes involving negative emotions but, as discussed at length, it can also be triggered by other cognitive processes, such as the perception of sharp contours and unusual stimuli. In addition, processes involving negative emotions could be implemented in other neural locations. Still, as long as we can order these links in a probabilistic way, and provided that we factor in the relevant task, neuroimaging data can be used to discriminate among competing cognitive hypotheses. Similarly, pattern-based inferences are also compatible with multiple realizability, even in its most radical forms. In the recognition example, multi-voxel patterns are extracted and classifiers are trained in specific tasks and for each subject individually. For instance, that some pattern is categorized as a recollection process by a classifier trained for a subject does not entail that the same pattern would be so categorized by a classifier trained on a different subject. Likewise, that a classifier trained for a subject in a particular recollection/familiarity task is reliable does not mean that it would still reliably distinguish between these processes in different tasks, e.g., one that uses visual objects instead of words. In short, the successful use of these patterns and classifiers to discriminate between theories of recognition does not depend on whether they are stable across subjects or even, within certain limits, across tasks. Hence, the assumption that recollection and familiarity processes are multiply realizable leaves the applicability of context-sensitive reverse inferences completely unscathed.

5.3 Revisiting Marr-Levels and Reductionism

Let us conclude by discussing the third and most general implication of our account. The classic reductive model of interlevel relations and Marr's influential subdivision of the study of cognition into three levels are, strictly speaking, independent. Early eliminative materialists such as Paul Churchland (1981) endorse reductionism while rejecting Marr-levels, whereas many philosophers recognize the usefulness of Marr-levels while eschewing reductionism (Bechtel and Mundale 1999). However, the two views mutually support each other. To wit, a standard reductionist response to multiple realizability is to argue that antireductionists set up a straw man by selecting relata on the cognitive side that are too coarse-grained to be reduced (Kim 1992; Shapiro 2000). The general idea underlying this response is that, as cognitive functions are progressively broken down into smaller subcomponents, it becomes more likely that we will reach a level where (local) reductive bridge laws can be established. Note how this picture of functional decompositions and local reductions fits in naturally with a standard interpretation of Marr-levels, according to which it only makes sense to ask about the lower-level implementation of functions once the cognitive processes that compute them have been laid out in algorithmic detail.

We do not deny that hypotheses regarding the neural implementation of cognitive processes constitute a significant portion of cognitive neuroscience. Indeed, astonishing progress has been made in the study of how certain perceptual and motor functions are carried out in the brain. However, we believe that this model of the psycho-neural interface as essentially addressing Marr-level 3 hypotheses is inadequate, as it leaves out much of the cognitive neuroscience of higher cognition. On the reductive account of Marr-levels, psychology and neuroscience only begin to meaningfully interact once we can ask how cognitive processes are implemented in neural hardware. This ignores a different—but equally important—type of psycho-neural interaction: using neural data to select among competing cognitive processes even when we have no clue how they could be neurally implemented (Del Pinal and Nathan 2013). This possibility of delving into the neural level only to 'come back up' to select hypotheses at the cognitive level is too often ignored by critics.

Our account of associative bridge laws also clarifies why, contrary to reductionist assumptions, it is often easier to employ neural data when Marr-level 2 hypotheses are not (yet) fully developed. For example, syntactic and semantic theories in linguistics are quite refined, but neuroimaging studies have been notoriously difficult to apply in this area. Linguists often face the task of determining whether a certain process is syntactic or semantic, with different models yielding different predictions. Take the case of 'it is raining,' used to mean that it is raining at the place of utterance. To account for this implicit location restriction, some models assume that a syntactic variable is inserted in the sentence prior to semantic interpretation (Stanley 2000); other models assume that the meaning of 'raining' is enriched to include the specification of a location (Recanati 2011). The former explanation appeals to a syntactic process; the latter to a semantic one. If we found bridge laws mapping syntactic and semantic

operations onto distinct locations or patterns of neural activation, we could try to discriminate between the two models by scanning subjects while processing such sentences. Unfortunately, establishing the relevant bridge laws is proving to be a daunting task: since semantic and syntactic processes usually work in tandem, they are extremely hard to disentangle. As a consequence, we cannot, at present, use neural data to discriminate between syntactic and semantic models of ellipsis. In contrast, models of moral and economic decision making are still comparatively underdeveloped. As Camerer and colleagues (2005) argue in great detail, one of the main divisions in current studies of decision making is between hypotheses that assume more rational processes, and hypotheses that assume an essential involvement of emotions. This division is illustrated by our discussion of moral decision making, and also emerges in several neuroeconomic debates, such as in competing explanations of the endowment effect (Knutson et al. 2008). This contrast is significant for the use of reverse inferences because we have bridge laws that map emotions and rule-guided behavior onto distinct brain regions (Miller and Cohen 2001; Greene 2009). Consequently, we can often test these decision-making hypotheses using reverse inference. However, as this branch of science progresses and mixed models that incorporate both rational and emotional components become more common, it may become more difficult to use our current bridge-laws to discriminate amongst them in neuroimaging studies.

The occasional difficulty in finding bridge laws that discriminate between advanced Marr-level 2 models, compared to the relative ease with which such laws often discriminate more elementary models, is hard to reconcile with the traditional reductive interpretation of Marr's framework. Hypotheses that have an advanced functional decomposition are better suited for implementation; hence, from the reductive perspective, they should also be better candidates for interaction and integration with the neural level. Furthermore, since few of our current hypotheses regarding capacities such as language or decision making are ready for Marr-level 3 implementation, it is hardly surprising that those who accept the reductive interpretation of Marr levels typically endorse the relative autonomy of the psychology of higher cognition. In contrast, our dynamic account makes better sense of the current limitations and achievements of interdisciplinary research at the border of psychology and neuroscience. Once again, our approach is compatible with the possibility that scientists will eventually discover the neural implementation of higher-level cognitive processes. Yet, abandoning the reductive perspective suggests other significant ways in which neural data can be employed to advance psychology.

References

Anderson, M. (2010). Review of *Neuroeconomics: Decision Making and the Brain*, eds. Glimcher, Camerer, Fehr, and Poldrack. *Journal of Economic Psychology 31*, 151–54.

Bechtel, W. and J. Mundale (1999). Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of Science 66*, 175–207.

Bickle, J. (1998). *Psychoneural Reduction: The New Wave*. Cambridge, MA: MIT Press.

Bickle, J. (2003). *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Dordrecht: Kluwer.

Bourget, D. and D. Chalmers (2013). What do philosophers believe? *Philosophical Studies*, 1–36.

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

Churchland, P. (1986). *Neurophilosophy*. Cambridge, MA: MIT Press.

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy 78*(2), 67–90.

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.

Enç, B. (1983). In defense of identity theory. *The Journal of Philosophy 80*, 279–298.

Fazekas, P. (2009). Reconsidering the role of bridge laws in inter-theoretic relations. *Erkenntnis 71*, 303–22.

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

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*.

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.

Gallistel, C. R. (2009). The neural mechanisms that underlie decision making. In P. W. Glimcher, C. F. Camerer, E. Fehr, and R. A. Poldrack (Eds.), *Neuroeconomics: Decision Theory and the Brain*, pp. 419–24. Elsevier.

Gazzaniga, M. S. (Ed.) (2009). *The Cognitive Neurosciences* (Fourth ed.). Cambridge, MA: MIT Press.

Glimcher, P. W. and E. Fehr (Eds.) (2014). *Neuroeconomics: Decision Making and the Brain* (2nd ed.). Burlington, MA: Elsevier.

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.

Harrison, S. A. and F. Tong (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature 458*, 632–35.

Henson, R. (2005). What Can Functional Neuroimaging Tell the Experimental Psychologist? *Quarterly Journal of Experimental Psychology 58A*, 193–233.

Hooker, C. A. (1981). Towards a general theory of reduction. Part III: Cross-categorical reductions. *Dialogue 20*, 496–529.

Horst, S. (2007). *Beyond Reduction: Philosophy of Mind and Post-Reductionist Philosophy of Science*. New York: Oxford University Press.

Hutzler, F. (2014). Reverse inference is not a fallacy per se: Cognitive processes can be inferred from functional imaging data. *Neuroimage 84*, 1061–69.

Joyce, J. (2011). A defense of imprecise credences in inference and decision making. In T. Szabo Gendler and J. Hawthorne (Eds.), *Oxford Studies in Epistemology*, Volume 4.

Kandel, E., J. Schwartz, T. Jessell, and S. Siegelbaum (2013). *Principles of Naural Science* (5th ed.). New York: McGraw-Hill.

Kim, J. (1992). Multiple realization and the metaphysics of reduction. *Philosophy and Phenomenological Research 52*, 1–26.

Kim, J. (1999). *Mind in a Physical World*. Cambridge, MA: MIT Press.

Kim, J. (2005). *Physicalism, Or Something Near Enough*. Princeton, NJ: Princeton University Press.

Kim, J. (2006). Emergence: Core ideas and issues. *Synthese 151*, 547–59.

Klein, C. (2009). Reduction without reductionism: A defense of Nagel on connectability. *The Philosophical Quarterly 59*(234), 39–53.

Klein, C. (2011). The dual track theory of moral decision-making: A critique of the neuroimaging evidence. *Neuroethics 4*, 143–62.

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.

Kuhn, T. S. (1977). Objectivity, value judgment, and theory choice. In *The Essential Tension: Selected Studies in Scientific Tradition and Change*, pp. 320–339. Chicago, IL: University of Chicago Press.

Legrenzi, P. and C. Umilta (2011). *Neuromania*. New York: Oxford University Press.

Lewis, D. K. (1969). Review of art, mind, and religion. *The Journal of Philosophy 66*, 23–35.

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.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman.

Marras, A. (2002). Kim on reduction. *Erkenntnis 57*, 231–57.

Mather, M., J. T. Cacioppo, and N. Kanwisher (Eds.) (2013). *20 Years of fMRI—What Has It Done for Understanding Cognition*, Volume 8. *Perspectives on Psychological Science*.

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.

Nagel, E. (1961). *The Structure of Science*. New York: Harcourt Brace.

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.

Norman, K., J. Quamme, and D. Weiss (2010). Listening for recollection: a multi-voxel pattern analysis of recognition memory retrieval strategies. *Frontiers in Human Neuroscience 4*, 1–12.

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.

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).

Putnam, H. (1967). Psychological predicates. In W. Capitan and D. Merrill (Eds.), *Art, Mind, and Religion*, pp. 37–48. Pittsburgh, PA: University of Pittsburgh Press.

Recanati, F. (2011). *Truth-Conditional Pragmatics*. Oxford: Oxford University Press.

Ross, D. (2008). Two styles of neuroeconomics. *Economics and Philosophy 24*, 473–83.

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

Shapiro, L. A. (2000). Multiple realizations. *The Journal of Philosophy 97*(12), 635–54.

Stanley, J. (2000). Context and logical form. *Linguistics and Philosophy 23*, 391–434.

Tong, F. and M. S. Pratte (2012). Decoding patterns of human brain activity. *Annual Review of Psychology 63*, 438–509.

Uttal, W. R. (2001). *The New Phrenology: The Limits of Localizing Cognitive Processes*. Cambridge, MA: MIT Press.

Uttal, W. R. (2002). Precis of the new phrenology: The limits of localizing cognitive processes in the brain. *Brain and Mind 3*(2), 221–28.

Van Orden, G. C. and K. R. Paap (1997). Functional Neuroimages Fail to Discover Pieces of Mind in the Parts of the Brain. *Philosophy of Science 64*, S85–94.