There and Up Again
On the Uses and Misuses of Neuroimaging in Psychology*†

Guillermo Del Pinal[*] and Marco J. Nathan[**]

[*]Columbia University
[**]University of Denver

November 14, 2013

Abstract

The aim of this article is to discuss the conditions under which functional neuroimaging can contribute to the study of higher-cognition. We begin by presenting two case studies—on moral and economic decision-making—which will help us identify and examine one of the main ways in which neuroimaging can help advance the study of higher cognition. We agree with critics that fMRI studies seldom 'refine' or 'confirm' particular psychological hypotheses, or even provide details of the neural implementation of cognitive functions. However, we suggest that neuroimaging can support psychology in a different way, namely, by selecting among competing hypotheses of the cognitive mechanisms underlying some mental function. One of the main ways in which neurimaging can be used for hypothesis selection is via reverse inferences, which we here examine in detail. Despite frequent claims to the contrary, we argue that successful reverse inferences do not assume any strong or objectionable form of reductionism or functional locationism. Moreover, our discussion illustrates that reverse inferences can be successful at early stages of psychological theorizing, when models of the cognitive mechanisms are only partially developed.

Keywords: functional neuroimaging; reverse inference; cognitive mechanisms; reductionism; decision-making; endowment-effect; higher-cognition.

*G. Del Pinal and M. J. Nathan contributed equally to this work.

†We are grateful to Russell Poldrack, Daniel Rothchild, anonymous referees and, especially, John Bickle for constructive comments on various versions of this essay. Some of the ideas developed here were presented at the European School of Molecular Medicine in Milan, Italy: the audience provided helpful feedback.

1 Introduction

In the entire decade of the 1980s, around a hundred books on cognitive neuroscience were published; in the following decade the number increased tenfold to 1,000; and since the turn of the century more than 10,000 volumes have been added, many of which were written for a lay audience (Aminoff 2009). This explosion of interest in cognitive neuroscience has been spurred by the development of important theoretical, experimental, and technological tools, and has given the general public the confidence that we are entering an era in which the deepest puzzles of the mind—related to consciousness, free-will, language, decision-making and other domains of 'higher-cognition'—will be finally unlocked. Such confidence is reinforced by the widespread practice among prominent psychologists and neuroscientists, especially evident in their popular works, to claim that some recent neuroscientific study *supports*, *confirms*, or *refines* some controversial psychological claim or theory (Hauser 2006; Iacoboni et al. 2007; Baron-Cohen 2011).

Still, not everyone is so enthusiastic: a substantial number of prominent psychologists and philosophers oppose this growing trend and argue that neuroscientific research—especially in the form of neuroimaging studies—has not improved our theories of higher-cognition (Uttal 2011; Satel and Lilienfeld 2013). Some authors claim that this is partly because our current technologies are too coarse and theories too undeveloped to begin to profitably investigate details about their neural implementation (Poeppel 1996; Coltheart 2004). Others are even more skeptical: they claim that even with improved technologies and more advanced theories, neuroscientific data *in principle* cannot help us advance the study of higher-cognition, either because of the supposed irreducibility or 'autonomy' of psychology, or because the current use of such data depends on fundamentally incorrect 'locationist' assumptions about the way in which the brain instantiates higher-level functions of the mind (Van Orden and Paap 1997; Fodor 1999; Uttal 2001). The significance of this discussion can hardly be overstated: the outcome of the debate about whether and in what ways neuroimaging—and, more generally, neuroscientific data—can advance our psychological theories of higher-cognition will partly determine how the study of the mind and brain will be approached and funded in coming decades.

The aim of this article is to discuss the conditions under which functional neuroimaging can contribute to the study of higher-cognition. We begin by presenting two case studies—one on moral and one on economic decision-making—which will help us identify and examine a specific way in which neuroimaging can advance psychological theories (Section 2). We agree with critics that fMRI studies seldom 'refine' or 'confirm' particular psychological hypotheses, or even provide details of the neural implementation of mental functions. However, building on insightful discussions in the literature (Poldrack and Wagner 2004; Henson 2005), we suggest that neuroimaging often supports psychology in a different way, namely, by selecting among competing hypotheses of the cognitive mechanisms underlying some mental function. We examine in detail one of the main ways in which neuroimaging can be used for hypothesis selection, namely,

reverse inferences in which the engagement of a cognitive state or process in some task is inferred from certain patterns or locations of neural activation. We argue that successful reverse inferences need not assume any strong or objectionable form of reductionism or functional locationism. Moreover, our discussion illustrates that neuroimaging can be used at early stages of psychological theorizing, when accounts of the cognitive mechanisms are only partially developed (Section 3).

2 Case Studies

The following case studies have a simple structure. In each case, we begin by specifying a widely accepted *psychological generalization*, call it 'G', about some domain of human decision-making. G is a Marr-level 1 hypothesis: it specifies a cognitive task or function which is performed by the human mind.[1] While it might be tempting to view neuroimaging as bearing directly on psychology by refining or (dis)confirming hypotheses at Marr-level 1, we argue that this is a mistake. We then introduce two competing hypotheses—M and M^*—about the *cognitive* mechanism(s) that underlie G. These are Marr-level 2 hypotheses: they specify the mechanisms that *compute* the Marr-Level 1 generalization G. It is at *this* level, we maintain, that neuroimaging can be used to advance theories of higher-cognition, mainly, by discriminating between competing cognitive mechanisms underlying Marr-level 1 generalizations. Marr-level 2 hypotheses such as M and M^* are purely *psychological* hypotheses; so for them to have identifiable implications at the neural level, at least some of their concepts or operations have to fall under the scope of *bridge-laws*—principles that associate psychological concepts or operations with patterns or locations of neural activation. We introduce the relevant bridge-laws in each of the case studies. Such laws raise a number of important questions, which can be better addressed *after* we examine their role in practice; for this reason, we postpone a discussion of their theoretical significance until Section 3.

[1] Marr (1982) famously argued that information-processing systems should be investigated at three distinct, complementary levels. Marr-level 1 hypotheses pose the computational problem: they state the task or function which the system performs or computes. Marr-level-2 hypotheses state the algorithm used to compute Marr-level 1 functions: they specify the basic representations and operations which the system uses to perform these functions. Finally, Marr-level 3 hypotheses 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 *physical* level. We should make clear that, in talking about Marr-levels, we are not committing to any rigid structural hierarchy or strict serial order of investigation. Our point is simply that one cannot profitably discuss (Marr-Level 2) algorithms or (Marr-level 3) issues about implementation, unless one has a general understanding of the (Marr-level 1) function(s) to be computed. As illustrated in the following case studies, once the Marr-level 1 function(s) have been laid out, investigations at all levels can process in parallel, mutually constraining each other in all directions (Henson 2005).

Case Study 1. Decision-Making in Moral Psychology: Trolley Problems

Our first case study focuses on a domain of higher-cognition that has recently gained the attention of philosophers, psychologists, and cognitive neuroscientists: moral judgments (**?**Sinnott-Armstrong 2008a,c). The generalization we consider is based on a widespread response to two well-known moral-choice scenarios (Thomson 1976). The first scenario (*switch*) asks you to imagine that a runaway trolley is about to run over and kill five people, but that you (and only you) can save them by hitting a switch that will divert the trolley onto a side track, in which case the trolley will run over and kill a single person. The second scenario (*footbridge*) asks you to imagine that a runaway trolley once again threatens five people but, in this case, the only way to save them is for you to throw a large person off a footbridge and into the trolley's path, which will stop the trolley but kill the person. Researchers asked subjects what they ought to do in each of these cases, and most subjects answer that you should hit the switch in *switch*, but that you should not throw the person in *footbridge* (Greene et al. 2001).

What is puzzling about this response is that, under most salient descriptions, the outcome of each choice is the same in both cases: one dead and five alive (press switch/push person) vs. one alive and five dead (do not press switch/do not push person). If the mechanism for moral decision-making followed consequentialist rules, then in both cases most respondents should choose the option that saves five people. Yet, repeated experiments have disconfirmed this prediction: most subjects say that you should press the switch, but that you should *not* throw the person. These patterns of moral decision-making are captured by the following generalization:

(G_T) In a situation where subject s faces the option to perform an action a that will result in the death of fewer people than would die if s were not to perform a, most subjects would choose a, unless doing so requires using a person directly as a means.

G_T is a testable descriptive (Marr-level 1) psychological generalization which applies to a certain restricted domain of moral decision-making. More specifically, G_T predicts that, when given two options, subjects will choose the option that is prescribed by consequentialist calculations, except when doing so involves using another person directly as a means, in which case subjects often refrain from performing that action.

The question we want to address is whether neuroimagining can be employed to advance psychological theories of moral decision-making. To answer this, we need to specify what would constitute an improvement of a psychological generalization. Below are three different kinds of advancements—often conflated in the literature—which correspond to three sorts of questions that can be raised about statements such as G_T.

(CONFIRMATION) The first obvious question is related to the confirmation of psychological generalizations. Is G_T a correct generalization of moral

decision-making? Are there other ways to test it, e.g., by comparing the responses to *switch* and *footbridge* to responses to different—yet related—moral scenarios?

(REFINEMENT) A different question is related to the descriptive accuracy of G_T. Can we express G_T in more precise terms, so that it better fits patterns of moral decision-making and makes predictions in cases not yet tested? One way to do this is to operationalize its vague concepts; for instance, one could specify what 'using someone directly as a means' actually amounts to.[2] Another option is to turn its qualitative distinctions into quantitative ones, e.g., roughly how many more lives need to be saved for most subjects to override their unwillingness to use a person directly as a means?

(EXPLANATION) A third type of question concerns the psychological mechanisms that generate the decision-making pattern captured by G_T. *Why* does something like G_T hold? In this case, we are looking for the internal cognitive structures and processes (at Marr-level 2) that determine subjects' response to moral scenarios such as *switch* and *footbridge*.

Answering any of these questions would constitute a substantial advancement over the original generalization, G_T. But all three questions contribute to the problem in very different ways and can be addressed using different experimental and theoretical tools. When scholars debate whether neuroimaging advances our theories of higher-cognition, it is often unclear which of the previous sorts of questions they have in mind and what kind of contributions their data are supposed to make to psychological theorizing.

Competing Psychological Explanations of G_T

The number of neuroscientific studies focusing on moral decision-making has been rising exponentially (Moll et al. 2008; Greene 2009). Interestingly, few—if any—of these works address questions related to the CONFIRMATION or REFINEMENT of Marr-level 1 generalizations. Most studies attempt to establish bridge-laws that map emotions which allegedly play a central role in moral decision-making—fear, distress, disgust, pain, anger, guilt, shame, etc.—to regions of the brain (Greene 2009; Phelps and Delgado 2009). Others try to establish bridge-laws that connect complex *cognitive capacities*, such as the ability to resolve cases of conflicting values or obligations, with their neural underpinnings (**?**). Only a few authors directly address questions about EXPLANATION,

[2] Greene et al. (2001, 2004a) propose some general conditions for distinguishing cases in which subjects believe they are using someone as a means ('personal' cases) from cases in which they believe that someone is injured as a result of collateral or unintended damage ('impersonal' cases). The difference is that, in personal cases, the harm is directly *authored* by the subject and affects a particular person, whereas in impersonal cases the subject is merely *editing* a pre-existing event. Whether this operationalization successfully captures the distinction at play is a controversial question, albeit one that transcends our present concerns (for discussion, see Mikhail (2008) and Greene (2008)). The important point is simply that in paradigmatic cases of not/using someone directly as a means something like G_T is correct.

i.e., use neuroimaging to discriminate between competing cognitive mechanisms (Trssoldi et al. 2012). However, we contend that this is one of the main ways in which neuroimagining can contribute to psychology.

To illustrate, consider again G_T, which systematizes a moral decision-making pattern in which subjects respond differently to *switch* and *footbridge* scenarios. G_T attributes this difference to the fact that in *footbridge*, but not in *switch*, subjects perceive their action as using another human being directly as a means. Let us assume that this generalization is essentially correct. We are then confronted with another question: *why* does perceiving an action as using another person directly as a means influence our moral decisions in that way? We can answer this question by proposing purely psychological Marr-level 2 hypotheses of the mechanisms underlying G_T. To wit, consider the following competing explanations, one involving mechanisms of *rational* cognition (M_R) and the other involving additional mechanisms of *emotional* cognition (M_E):[3]

(M_R) Moral decision-making consists in applying abstract rules. One (consequentialist) rule is to maximize utility, but this principle is sometimes blocked by other abstract rules, one of which says that people cannot be used directly as means (a deontological rule). In other words, perception of 'direct sacrifice' activates a deontological rule which blocks utilitarian calculations in cases like *footbridge*. This is why the consequentialist rule is followed in *switch* but not in *footbridge*.

(M_E) Moral decision-making *partly* consists in applying abstract rules. One rule is to maximize utility, but this rule is sometimes blocked or disengaged by certain negative emotions—such as fear, distress, or disgust—that are triggered by certain cues that signal that we are directly using someone as a means. In other words, perception of 'direct sacrifice' generates negative *emotions* that block utilitarian calculations in cases like *footbridge*. This is why the consequentialist rule is followed in *switch* but not in *footbridge*.

Despite structural similarities, M_R and M_E are very different explanations of G_T. M_R accounts for the different response in *switch* and *footbridge* by postulating a conflict between rules (consequentialist vs. deontological). In contrast, M_E explains the difference by postulating a conflict between (consequentialist) rules and negative emotions.

How should one decide between M_R and M_E? We maintain that neuroimaging can provide crucial evidence, favoring M_E. Specifically, fMRI studies suggest that what blocks the utilitarian calculation in cases such as *footbridge* is a conflicting negative emotion rather than a deontological rule. If correct, this supports our thesis about how neuroimaging advances psychological theorizing: when two competing Marr-level 2 hypotheses are pitted against each other, we can often devise fMRI experiments that unambiguously favor one of them.

[3]Of course, M_R and M_E are not the *only* plausible explanations of G_T. Our claim is simply that they are two reasonable, competing hypotheses, close versions of which have been defended by philosophers and psychologists (Kohlberg 1971). For recent discussion of rationalism in moral psychology, see Joyce (2008).

Neuroimaging Evidence

Given any purely psychological mechanism M, in order to determine N_M—the neural correlate of M or of one of its subcomponents—one needs to appeal to bridge-laws. Without these laws, one cannot link psychological concepts or mechanisms to their underlying patterns or locations of neural activation, in which case neuroscientific data cannot be brought to bear on competing psychological mechanisms. This is why neuroimaging studies seeking to establish such correlations, although often disparaged by critics (Fodor 1999), are an integral part of the more general effort to advance psychology.

Consider, once again, the competing explanations of G_T: M_R and M_E. According to M_R, the mechanism that accounts for G_T is based on consequentialist and deontological rules. The consequentialist rule tells subjects to select the option which saves the most lives, which explains why they would flip the switch. The deontological rule restricts the domain of the consequentialist one by stating that you cannot use someone directly as a means, which explains why most subjects would refuse to push the person. M_R allows that these choices generate negative emotional reactions, perhaps as side-effects. However, according to M_R, these negative emotions are not part of the mechanism that determines the usual response in cases like *footbridge*. In contrast, according to M_E, what interferes with the consequentialist rule is the involvement of certain negative emotions triggered by the perception of using someone directly as a means. Hence, in cases like *footbridge*, M_E and M_R make different predictions. M_R requires differential activation in areas of the brain associated with abstract reasoning and rule application, and it is compatible—but does not require—activation in areas of the brain associated with negative emotions. M_E makes the opposite prediction: it requires differential activation in areas associated with negative emotions, and it is compatible—but does not require—activation in areas associated with abstract reasoning and rule application.

These predictions have been tested. In a famous study, Greene et al. (2001) used fMRI to scan subjects while they responded to *switch*, *footbridge*, and similar scenarios. As expected, most subjects reported that they would flip the switch but would not push the person. Greene and colleagues discovered that *footbridge*-like scenarios, compared to *switch*, differentially activated brain regions associated with negative emotions—such as the medial prefrontal cortex, the posterior cingulate cortex, and the amygdala. In contrast, *switch*-like scenarios, compared to *footbridge*, differentially activated brain regions of the dorsolateral prefrontal cortex associated with working memory (Smith and Jonides 1997) and cognitive control (Miller and Cohen 2001). These results support M_E over M_R, for, although M_R is *compatible* with emotional associations in *footbridge*, it predicts that the same kind of rule-based reasoning goes on in both kinds of scenarios, the difference in outcome resulting from the application of different rules. However, the neuroimaging data is in tension with that prediction: areas associated with rule-based reasoning and conflict-resolution were not differentially activated in *footbridge*, relative to their level of activation in *switch* and other non-moral control scenarios that required rule-application.

We should note that, just like any other experiments that aim to discriminate between competing psychological hypotheses, these fMRI studies are neither intended to be definitive nor to be considered in isolation. As emphasized by Greene (2008, 2009), these results should be seen as part of a nexus of behavioral and neuroscientific studies, which together support M_E over M_R. For example, Valdesolo and DeSteno (2005) found that normal subjects are more likely to push the person in *footbridge* following a positive emotion induction aimed at counteracting negative emotional response. In addition, patients affected by lesions in the ventromedial prefrontal cortex, which is thought to work in concert with the amygdala (Schoenbaum and Roesch 2005), or affected by frontotemporal dementia, who have emotional deficits but normal abstract-reasoning capacities, were disproportionately likely to approve pushing the person (Mendez et al. 2005; Koenings et al. 2007; Ciaramelli et al. 2007). These studies suggest that the role of negative emotions in the normal response to *footbridge* is *causal*, as predicted by M_E, and not merely an associated side-effect, as predicted by M_R.

Case Study 2. Decision-Making in Behavioral Economics: The Endowment Effect

We now shift to a different domain—*economic decision-making*—which provides us with another example of the way in which neuroimaging can bear on competing theories at Marr-level 2. According to rational choice theory, valuation is *reference independent*, in the sense that the value that subjects assign to goods should not vary relative to their allocation. A corollary of this thesis is that ownership should not influence preferences: if subject s assigns value v to good g, then v should not vary as a function of whether g is owned by s or by someone else. However, decades of behavioral research have shown that this prediction is systematically violated: even when allocation occurs randomly, people consistently prefer goods they own to similar goods that they do not own.

In a celebrated study, Kahneman et al. (1990b) gave a group of participants (the 'sellers') a coffee mug with a university logo and told them the mug was theirs to have. The experimenters then presented them with a series of prices ranging from \$0.25 to \$9.25 and for each price asked them if they would be willing to sell the mug. A second group (the 'buyers') were asked if they would be willing to buy a mug for each price from the same set. Finally, a third group (the 'choosers') were asked whether they would prefer the mug or the money for each price of the same set. Note that the choosers and sellers were given identical options (walk away with mug or money); hence, according to reference-independent rational choice theory, their choices should be roughly the same. However, this prediction was systematically violated: sellers quoted a median price of \$7.12, while choosers quoted a median price of \$3.12, only slightly higher than buyers, who quoted a median price of \$2.87.

This pattern—where the price that subjects are willing to accept to part with a good exceeds the price that they are willing to pay to acquire the same good—

is a robust psychological generalization, known as the *Endowment Effect*, which can be expressed as follows:

(G_E) Assume that $g_1 \dots g_n$ are goods; V_s is a function that assigns to any g_i s's selling value (the minimum value at which s would be willing to sell g_i); V'_s is a function that assigns to any g_i s's buying value (the maximum value at which s would be willing to buy g_i). Owning a good tends to increase its value in the following way:

$$V_s(g_i) = a(V'_s(g_i)), \text{ where } 1.2 < a < 2.5$$

G_E is a Marr-level 1 psychological generalization, which applies to a domain of economic decision-making. Although it is now widely accepted by both economists and psychologists, G_E is by no means a trivial claim: it violates reference-independent rational choice theory—part of the foundations of classical economics—which (incorrectly) predicts that $V_s(g_i) \approx V'_s(g_i)$.

Just as in Case Study 1, we can raise (at least) three different types of questions about G_E. Answering any of these questions would constitute a substantial advancement in our theories of economic decision-making.

(CONFIRMATION) The first question concerns the confirmation of psychological generalizations. In this particular case, G_E hardly requires any additional confirmation; however, this is the exception rather than the rule. For most Marr-level 1 hypotheses we can ask: what kind of further evidence do we need to establish that the hypothesis is reasonably accurate?

(REFINEMENT) How can we improve the descriptive accuracy of G_E? For example, is it possible to express G_E in more precise terms, so that it better fits the data and entails more testable predictions? One option is to determine the size of a as a function of types of goods (e.g., luxury vs. common goods) or the duration and origin of ownership. Another option is to specify in greater detail the 'boundaries' of G_E, e.g., the conditions under which a good becomes a good for use (which falls under G_E) vs. a good for exchange (which does not fall under G_E).

(EXPLANATION) The third question concerns the psychological mechanism(s) that underlie G_E—*why* does G_E hold? Finding such cognitive processes is especially pressing when the evidence for the generalization is such that there can hardly be any doubt that it captures a relevant mental function, as in the case of G_E.

As emphasized in Case Study 1, when scientists and philosophers ask whether neuroimaging can advance psychological theorizing, they often fail to clarify which of these very different sorts of questions they have in mind.

Competing Psychological Explanations of G_E

In the case of G_E, neuroimaging has not contributed to psychology by addressing questions about either REFINEMENT or CONFIRMATION. This becomes obvious once we realize that G_E was postulated, refined, and confirmed using only

psychological concepts and behavioral experiments. However, just like in Case Study 1, neuroimaging can help us advance questions about EXPLANATION, i.e., about Marr-level 2 psychological mechanism(s) underlying G_E.

There are various competing explanations of the mechanisms underlying G_E (Novemsky and Kahneman 2005; Ariely et al. 2005; Nayakankuppam and Mishra 2005; Rick 2011). Here we only consider two amongst the most influential, mainly because these are the ones that have been submitted to fMRI investigation (Knutson et al. 2008).[4] Just as in the previous case study, one of the hypotheses is more emotional, while the other is more cognitive. The first theory, M_L, relies on a particular account of the mechanisms underlying *loss-aversion* which appeals to negative emotions. The other theory, M_P, relies instead on certain cognitive mechanisms of differential *perception*:

(M_L) G_E is a consequence of fear/distress-induced *loss-aversion*—the tendency for losses to have greater hedonic impact than comparable gains. Goods-to-buy are typically treated as gains; but ownership resets the reference point so that goods-to-sell are treated as losses (Kahneman et al. 1990b,a). When goods are considered as potential losses, their negative hedonic impact is 'exaggerated' due to an emotional over-reaction of fear/distress at the prospect of loosing the good (Caramer 2005).[5]

(M_P) G_E is a consequence of *differential perception* between sellers and buyers. This is also a reference-dependent theory, but one which posits a different mechanism to determine the subjective value of goods: ownership causes subjects to focus more on the positive features and less on the negative features of goods, relative to buyers (Nayakankuppam and Mishra 2005).

M_L and M_P both assume that the mechanism(s) underlying choice include a subcomponent that assigns value to each option, but they differ in how this value assignment is determined. According to M_P, it is determined, in part, by the perspective taken on the relevant goods. This focus is more positive when goods are owned compared to when they are not; put informally, ownership causes goods to look better. In contrast, according to M_L, values assigned to owned goods are determined by an emotional 'over-reaction' that underlies loss-aversion: the prospect of parting with owned goods triggers a substantial degree of distress. We now illustrate how fMRI data can be used to choose between

[4]Other competing explanations of G_E include the claims that subjects over-value items they own because they are associated with the self (Morewedge et al. 2009), and that G_E is caused by the desire to avoid a bad deal (Brown 2005). For an overview see Rick (2011).

[5]Note that M_L is not equivalent—although it is intimately related—to 'loss aversion' as used in behavioral economics, where it usually refers to a Marr-level 1 generalization, covering both risky and riskless choices, according to which "changes for the worse (losses) loom larger than changes for the better" (Kahneman and Novemsky 2005, 119). Just as there is a debate about what mechanism explains G_E—or as it is sometimes called 'loss aversion in riskless choice'—there is also a debate about what explains loss aversion in general, when used to refer to a Marr-level 1 generalization. M_L is one such proposal, an early version of which was suggested by Caramer (2005). Sometimes theorists talk as if loss aversion 'explains' G_E; however, in the way we are using the term, G_E is only one particular manifestation of loss aversion, and the 'explanations' are Marr-level 2 claims such as M_L and M_P.

these competing explanations (cognitive vs. emotional) of the mechanism(s) underlying G_E, strongly favoring M_L over M_P.

Neuroimaging Evidence

The relevant event-related fMRI study is presented in Knutson et al. (2008). Using high-value consumer goods such as iPods and digital cameras, Knutson and colleagues elicited an endowment effect by asking subjects to purchase goods (with money they were given at the beginning of the session), sell other goods (also given to them at the beginning of the session), and choose between goods and money. During the *selling trials*, subjects were shown a good they had been given, offered a certain price, and then asked whether they wanted to keep the good or sell it. During the *buying trials*, subjects were shown a good, shown a certain price, and then asked whether they wanted to buy the good at that price. Finally, in the *choosing trials*, subjects were shown a good, then a price, and then asked to choose between the good and the money. All subjects engaged in each of the three tasks. During the trials, subjects were scanned with fMRI to determine and compare the areas of differential activation when viewing products in buying vs. selling vs. choosing conditions. The results were consistent with G_E: for each good, selling prices were significantly greater than choice prices, and choice prices were only marginally larger than buying prices.

In order to use the resulting fMRI data to bear on M_L and M_P, we have to appeal to the relevant bridge-laws. Knutson and colleagues make the following suggestion. They argue that if the mechanism underlying G_E was close to M_P then, while viewing products in the selling-condition relative to the buying-condition—the case in which products 'look' better—subjects should show increased differential activation in the nucleus accumbens (NAcc). The reason for this is that NAcc activation correlates with product preference and increased attraction to products, and is a reliable predictor of decisions to purchase (Knutson and Greer 2008). In contrast, if the mechanism underlying G_E were M_L then, while viewing items in the selling-condition relative to the buying-condition, subjects should display increased differential activation in the insula. Again, the rationale is straightforward. Insula activation has been associated with the anticipation and experience of distress (Sanfey et al. 2003; Eisenberger et al. 2003), and correlates positively with subjective measures of distress (Masten et al. 2009).

Let us consider the results of the experiment against this theoretical background. When viewing products in the sell-condition compared to the buy-condition, subjects showed increase right insula activation, which positively correlates with the size of the endowment effect (i.e., in our formulation, with the size of a in G_E). This suggests that the more distress subjects feel when contemplating parting with items they own, the more pronounced their endowment effect. In addition, increased NAcc activation, which correlates with increased attraction to products and predicts product-preference, was not observed in subjects in the sell-compared to the buy-condition, and hence did not correlate with the size of the endowment effect. These results suggest that the mecha-

nism responsible for G_E is closer to M_L than to M_P.[6] In short, the price at which subjects are willing to sell goods is higher than the price at which they are willing to buy them because the thought of giving up goods evokes negative emotions, and not because goods looks better or more attractive when one owns them.

Again, this study does not, by itself, conclusively settle the debate between M_L and M_P; it only substantially increases our confidence in M_L, especially when considered as part of a nexus of other neuroscientific and behavioral studies. Further support for M_L over M_P comes from experiments that elicited instances of G_E while also measuring the subjective attractiveness of goods. Under these circumstances, subjects endowed with particular goods did not rate them as more valuable or attractive than subjects not endowed with the same goods (Kahneman et al. 1990b,a). With some effort, each of these studies can be made compatible with M_P; yet, taken together, they provide converging evidence for M_L.

3 Reverse Arguments and Bridge-Laws

Critics of the idea that neuroimaging can contribute to the advancement of psychological theories of higher-cognition often talk as if the contribution of neuroimaging is limited to two projects (Fodor 1999):

(i) To show, via correlational studies, that every mental event or process has an underlying neural implementation.

(ii) To provide data that can be used to form hypotheses about how cognitive mechanisms are implemented in neural hardware.

Now, (i) is not a very interesting project, since no one seriously doubts the underlying claim; and (ii) is not a very realistic project, since no one really thinks that, presently, we can understand the details of how mechanisms of higher-cognition are implemented in the brain. However, our case studies suggest that neuroimaging can contribute in a way that is more interesting than (i) and a more realistic than (ii), namely, by discriminating between competing Marr-level 2 hypotheses. These case studies center on what is usually called *reverse inference* (Poldrack 2006) or *structure-to-function deduction* (Henson 2005), where the engagement of a cognitive process in a given task is inferred from the activation of a particular brain region. Let us call arguments based on reverse

[6] As in the previous case study, the more cognitive mechanism, M_P, is *compatible* with activation in areas associated with negative emotions. For example, one could suggest that, as a consequence of the positive perception triggered by ownership, subjects feel some distress at the prospect of loosing owned goods. What is important is that according to M_P, negative emotions are at best a consequence—not a subcomponent—of the mechanism which assigns value to the prospect of loosing an owned good. For a detailed discussion of why this compatibility does not undermine the suggestion that neuroimaging evidence favors M_L over M_P, see Section 3 below.

inferences, *reverse arguments*. We now turn to some general methodological issues raised by reverse arguments.

The basic form of reverse arguments is the following. In each case, we start out with the question of which among two competing cognitive mechanisms M or M^* are engaged in task Z, which is an instance of some important psychological generalization G. Suppose that M is the mechanism supported by the evidence, then the reverse argument has the following form:

(Premise 1) M is partly constituted by cognitive subprocesses or states $m_1, \ldots, m_n$; M^* is partly constituted by cognitive subprocesses or states $m_1^*, \ldots, m_n^*$.

(Premise 2) $m_1, \ldots, m_n$ are associated with activation in brain regions $n_1, \ldots, n_n$ via bridge-laws $Br_1(m), \ldots, Br_n(m_n)$. $m_1^*, \ldots, m_n^*$ are associated with activation in brain region $n_1^* \ldots, n_n^*$ via bridge-laws $Br_1^*(m_1^*) \ldots, Br_n^*(m_n^*)$.

(Premise 3) In task Z either (i) there was differential activation in $n_1, \ldots, n_n$ and not for at least one n_i^* that is not equal to any of $n_1, \ldots, n_n$; or (ii) there was differential activation in $n_1, \ldots, n_n$ and at least one n_i is not equal to any of $n_1^* \ldots, n_n^*$.

(Conclusion) Premises 2 and 3 entail that, in task Z, $m_1, \ldots, m_n$ are more likely engaged than $m_1^*, \ldots, m_n^*$. This result, together with Premise 1, entails that, in task Z, M is more likely engaged than M^*.

The key to reverse arguments lies in the nature and role of the bridge-laws used in their reverse inferences. We now raise four crucial points about bridge-laws that address some common objections against this type of argument.

1. Theorists often talk as if what bridge-laws typically associate with certain brain regions are *entire* Marr-level 2 cognitive mechanisms. However, this is not the strategy pursued by some of the best recent examples of neuroimaging studies of higher-cognition.

As Premises 1 and 2 make clear, reverse arguments typically subdivide the competing cognitive mechanisms into their parts, and strategically test the engagement of those components for which we have adequate bridge-laws. In other words, cognitive subprocesses or states $m_1, \ldots, m_n$ and $m_1^*, \ldots, m_n^*$ are not necessarily a *complete* analysis or breakdown of M and M^* respectively. Rather, they only stand for some important *subcomponents* of M and $M*$, mainly, the ones that fall under bridge-laws. For example, in the case of M_R vs. M_E, we are not concerned with whether *footbridge* triggers differential activation in areas associated with specifically deontological rule-application. We could pursue this strategy *if* we had bridge-laws to distinguish areas associated with deontological rule-application from those associated with consequentialist rule-application; but we currently have no such principles. Instead, Greene et al. (2001) wisely focus on areas associated with negative emotions for, on the one hand, we have bridge-laws that cover these sorts of emotions and, on the other hand, only one of the competing theories (M_E) requires essential involvement of negative

emotions in the mechanism that determines the judgments in *footbridge*. An analogous reason also explains why Greene and colleagues focus on areas associated with general rule-application (but not specifically with deontological vs consequentialist rules): this is because only one of the theories (M_R) predicts—mistakenly in this case—that the judgments in *footbridge* essentially involve areas associated with general rule-application.[7]

This point bears on two objections often raised in the literature, which we now address in turn. According to the first objection, to hold that neuroimaging can be used to advance theories of higher-cognition, one must assume that the mechanisms computing psychological functions are *localized* (Uttal 2002). Assuming the locality of cognitive mechanisms is plausible in the case of perceptual functions, the objection runs, but it is implausible in the case of arguably non-modular central processes, such as those that presumably compute the higher-cognition functions. If correct, this entails that the previous case studies implicitly—and implausibly—assume that the implementations of entire moral (M_R and M_E) and economic (M_L and M_P) decision-making mechanisms are localized in some fairly substantial sense. It should now be obvious why this objection is misguided. For the reverse argument to work, one only has to assume the locality of *some* key subcomponents of the competing decision-making mechanisms. As long as one of mechanisms, say M, has a particular subcomponent m_i that M^* does not have, and provided that we have a bridge-law that covers m_i, we can then look for the neural activation associated with m_i.[8] This, of course, does not mean that the M-hypothesis is entirely necessarily correct—that M is the true mechanism implemented by the brain to perform task Z. All we have shown is that, given the current evidence, M is more likely than M^*.

The second complaint is that the reasoning used in reverse arguments is *circular* (Van Orden and Paap 1997). The objection runs as follows. Studies that aim to establish bridge-laws infer that a particular brain region is involved in a cognitive state or process M by *assuming* that the cognitive process is engaged in some task Z. Then, reverse-argument studies test whether cognitive process M or M^* is involved in some task which is similar to Z. But given that we previously *assumed* that m is engaged in Z, and that the new task is similar to Z, it can hardly be surprising that we usually end up with a result that favors M over M^*. Our response is that this alleged circularity of reverse inferences

[7]Note that M is supported if *either* condition (i) or condition (ii) of Premise 3 is satisfied. However, the best reverse arguments try to fulfill both conditions. This is indeed the strategy adopted in Case Studies 1 and 2. To wit, in *footbridge* M_E predicts differential activation in areas associated with negative emotions, and M_R predicts differential activation in areas associated with rule-application. In the experiment, the former areas are activated—satisfying condition (ii)—and the latter areas are not—satisfying condition (i). To see this clearly, just assume that $M = M_E$; $M^* = M_R$; $m_1, \ldots, m_n$ = fear/distress $m_1^*, \ldots, m_n^*$ = consequentialist/deontological rule application; $n_1, \ldots, n_n$ = +posterior cingulate cortex, +amygdala; and that $n_1^* \ldots, n_n^*$ = +dorsolateral prefrontal cortex.

[8]Indeed, even if M and M^* share *all* subcomponents, we can often still test for patterns of neural activation that differentiate them, as long as their components are ordered differently. But in these cases it is usually more appropriate to employ a different type of inference, based on patterns rather than locations of activation. For further discussion see Henson (2006), Kriegeskorte (2011) and Section 4 below.

stems from an inaccurate description of reverse arguments. As illustrated by the previous case studies, in most neuroimaging experiments of higher-cognition what is tested is neither M nor M^* in their entirety, but rather some of their key subcomponent(s), $m_1, \ldots, m_n$ and $m_1*, \ldots, m_n*$ respectively. For example, in Case Study 1 we considered neural areas associated with controlled rule-application and areas associated with negative emotions. The bridge-laws for these subprocesses can be established using simple tasks that do not involve moral decision-making, e.g., by considering rule-following in nonmoral cases and by inducing fear, disgust, or distress *via* direct perceptual cues. Furthermore, the simple tasks used to establish bridge-laws are chosen precisely because they are cases in which we can safely assume the involvement of the cognitive process and states of interest.

2. There is a broad but important distinction between two types of bridge-laws often conflated in the literature: *reductionist* and *associationist* bridge-laws. Despite accusations of implicit reductionism, only associationist bridge-laws are needed to support reverse arguments.

What we call *reductionist* bridge-laws are law-like principles that purport to reduce—thereby refining or replacing—'upper-level' (cognitive) concepts or operations to 'lower-level' (neural) ones.[9] Many theorists correctly believe that reductionist bridge-laws are amongst the hardest to find in science, and that they can only be established when the variable whose identity is sought is "already well defined and understood in its own terms, at its own level of analysis, within a theoretical framework that rests on observations at that level" (Gallistel 2009, 421). While this is a sound methodological observation, it is a common mistake to assume that reductionist bridge-laws are *required* by reverse inferences. This mistake is partly responsible for the negative attitude towards reverse arguments

The bridge-laws used in Case Study 1 and 2, as well as in other studies which employ reverse inference to discriminate amongst competing cognitive mechanisms, are *not* reductionist bridge-laws. Rather, they are what we can call *associationist* bridge-laws, i.e. laws that associate—but do not refine or replace—cognitive concepts or operations with neural activation at certain locations. What associationist laws establish is simply that a certain psychological operation or state m_i is implemented in a specific neural location n_i. Thus, if m_i figures in only one of two competing mechanism underlying some cognitive task Z, then we can use these bridge-laws to support the claim that M (as opposed to $M*$) is engaged in Z. One way to understand the difference between associationist and reductionist bridge-laws is to consider their roles in two different projects. One project addresses *how* cognitive mechanisms are implemented in the brain. This task *is* an exercise in reductionism which requires reductionist

[9] Two examples of reductionist bridge-principles are laws that link the behavioral operation of nerve impulses to the electrophysiological process of action potentials, and (more controversially) laws that link cognitive measures of subjective utility to firing rates of groups of neurons in the frontal cortex and basal ganglia (Glimcher et al. 2009).

bridge-laws. A different project concerns *which* of two competing mechanisms is implemented by the brain. This is *not* an exercise in reductionism—since the competing mechanism are still at the cognitive level—and only requires associationist bridge-laws. The distinction between reductionist and associationist bridge-laws bears on the accusation, repeated *ad nauseam* in the literature, that using neuroimaging to advance theories of higher-cognition is tantamount to embracing some form of 'reductionism' (Fodor 1999; Uttal 2002; Ross 2008; Gallistel 2009). This charge misses the target. On the approach defended here, reverse arguments only require perspicuous associationist—but not necessarily reductionist—bridge laws.

3. The success of a reverse argument depends on the selectivity of the relevant brain regions. The selectivity of a brain region is inversely proportional to the number of bridge-laws that apply to it. However, the degree of confidence that a particular bridge-law applies in a given reverse argument is determined against the relevant tasks. This has been either ignored or misapplied in several recent critical discussions.

Following Poldrack (2006), several authors correctly emphasize that the success of a reverse argument partly depends on the degree of 'selectivity' of the relevant brain regions (Henson 2005; Phelps 2009). *If* the brain region in question, say n_1, only activates for the cognitive process of interest, say m_1, then the reverse inference from n_1 to m_1 is valid. However, most theorists agree that we presently do not have reason to hold that brain regions are maximally selective in that way: most brain regions, as currently individuated, are covered by sets of bridge-laws which associate them with various cognitive functions. In most cases, our confidence that a particular bridge-law applies is a matter of degrees, which is determined, as Poldrack suggests, by the conditional probability that cognitive state or process m_1 is engaged given activation in n_1:

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

Note that the prior $P(m_1)$ is conditioned on the task used in the reverse argument—in this case Z.[10] Importantly, (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, i.e., $P(n_1|\neg m_1)$.

[10] For readers unfamiliar with Bayesian formalism, note that Equation (1) is an application of Bayes' theorem, which tells you how to determine the conditional probability of hypothesis h given evidence e (intuitively, how to update your degree of belief in a hypothesis given new evidence):

$$P(h|e) = \frac{P(e|h)P(h)}{P(e)} = \frac{P(e|h)P(h)}{P(e|h)P(h) + P(e|\neg h)P(\neg h)}$$

Accordingly, (1) states that the probability that cognitive state m_1 is engaged, given n_1 (location or pattern of neural activation), is obtained by multiplying the probability that n_1 obtains given that cognitive state m_1 is engaged, by the prior probability of engagement of m_1, and dividing this value by the overall probability of n_1.

Critics argue that once it is clear that our confidence that a given bridge-law applies in some task is determined by (1), the reverse inferences underlying most 'interesting' reverse arguments turn out to be unacceptably weak (Miller 2008; Phelps 2009; Legrenzi and Umilta 2011). In some cases, such accusations are justified. However, the presumed lack of selectivity of brain regions which play a key role in neuroimaging studies of decision-making—e.g., the ventral striatum, the amygdala, and the insula—has been substantially exaggerated. The main reason for this is that, in the wake of Poldrack's (2006) influential discussion, (1) has been misused in two important ways.

The first problem is that we lack a reliable way of determining the crucial values of (1), especially $P(n_1|\neg m_1)$, often called the 'false alarm' rate. This value is determined by instances in which activation was observed in the absence of the cognitive process of interest. Part of the reason why we usually cannot determine the false alarm rate is that neuroimaging databases are not organized in the required way. To illustrate, consider how Poldrack (2006) uses the BrainMap database (http://www.brain-map.org/) to examine the strength of the reverse inference that activation in Broca's area implies engagement of language function. He found that, as of September 2005, there were 3,222 experimental comparisons, out of which there were (i) 166 language tasks in which Broca's area was differentially activated, (ii) 199 non-language tasks in which Broca's area was activated, (iii) 703 language tasks in which Broca's area was not activated, and (iv) 2154 non-language tasks in which Broca's area was not activated. Based on these numbers and assuming that the prior probability, $P(language)$, that a language function is engaged in some arbitrary task is 0.5, Poldrack concludes that $P(language|+Broca) = 0.69$. This results in a Bayes' factor of 2.3, which is generally considered a weak increase in confidence over the prior (Jeffreys 1961). The problem with Poldrack's calculation, however, is that we do not know which of the 3,222 experimental tasks are duplicates or involve similar tasks. Different experiments that duplicate the same task, or involve a similar one, should not count as independent evidence of the process-(un)specificity of a neural area. Hence, databases such as BrainMap which do not allow one to factor out those numbers are unreliable determinants of the degree of selectivity of a neural area.

The second way in which (1) has been misused by critics of reverse inference is the following. In his influential discussion, Poldrack (2006) emphasizes that the prior $P(m_1)$—and, consequently, $P(\neg m_1)$—is determined relative to a task Z; this means, of course, that Z also matters for determining $P(n_1|\neg m_1)$. After noting this, Poldrack set it aside to avoid some unnecessary complications. This had the unfortunate consequence that various critics of reverse inference have since overlooked this important task-relativity. To illustrate, let us consider an example related to Case Study 1: the selectivity of the amygdala, which is often employed in reverse inferences, especially in the work of neuroeconomists and moral psychologists. Phelps (2009) correctly points out that, although the amygdala is typically engaged in processes involving fear and other negative emotions, it is also involved in many other cognitive processes, typically unmentioned in studies such as Greene et al. (2001). Such processes include the per-

ception of odor intensity, sexually arousing stimuli, and trust from faces (Phelps 2006; Lindquist et al. 2012), as well as processing faces from other races and the perception of biological motion and sharp contours (Phelps 2009). Indeed, a new hypothesis has been recently advanced, according to which 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 argues that amygdala activation in a given task Z could mean that any of these other cognitive processes is engaged, which seems to imply that, independently of the nature of Z, the value $P(+amygdala|\neg fear)$ is considerably *higher* than usually assumed. This entails that the value of $P(fear|+amygdala)$—which is inversely proportional to $P(+amygdala|\neg fear)$—is considerably *lower* than usually assumed.

This suggestion, however, is misleading, for when we consider the *particular* task relative to which the value of $P(fear|+amygdala)$ is determined, we can see that, in most cases, the value of $P(+amygdala|\neg fear)$ is usually much lower than suggested by Phelps. For example, Case Study 1 involves the reverse inference from *differential* activation of the amygdala in *footbridge* cases relative to *switch* cases, to the engagement of processes involving negative emotions. Note that most of the cognitive processes which the amygdala is thought to also implement are not plausible candidates for differential engagement in *footbridge* relative to *switch*. To wit, relative to *switch*, *footbridge* does not differ in the presence of (or stimuli directly related to) odors, faces, sexuality, or sharp contours; and it is very plausible that, for most participants, *switch* is as unusual and novel as *footbridge*.[11] Now, the precise value of $P(fear|+amygdala)$ is impossible to calculate in any meaningful way, so the overall persuasiveness of the reverse argument presented in Case Study 1 is conditional on what future research will tell us about the selectivity of the relevant brain areas. Still, for the reasons just given, we can be more optimistic than Phelps and other critics about the selectivity of the relevant brain areas, once the sets of tasks relevant to particular reverse arguments are taken into account.

[11] John Bickle (p.c.) suggests a way in which this last point about novelty might be questioned. While utilitarian ethical dilemmas are rather commonplace, it is somewhat rare—and, hence, more novel—to be asked, as one is in *footbridge*, to explicitly use another person as the instrument triggering the action with the best overall consequences. In response, one can argue that what matters most for the degree of novelty of a stimulus are its fine grained particular (esp. perceptual) properties, and not whether at some abstract level it falls or does not fall under a type tokens of which are either novel or common. In this sense, being confronted with an event of pressing a switch in a trolley problem to save more people than would otherwise die is, for most respondents, as novel an event as being asked to throw someone off a bridge for the good outcome. If so, the hypothesis that the amygdala is activated due to the novelty of the *footbridge* tasks is still in tension with the observed differential activation in *footbridge* relative to the also novel *switch* tasks. However, even if this response is rejected, Bickle's comment allows to make a more important point. This is that once we reduce the number of cognitive processes associated with the activation of a certain brain area (in this case, the amygdala) by taking into account the relevant tasks (i.e, once we reduce the set of relevant bridge-laws), we can often directly control for the remaining possibilities. In our example, this could be done by altering the novelty of different tokens of *switch*-like and *footbridge*-like tasks.

We are not suggesting, of course, that taking into account the task-relativity of reverse arguments will in the end vindicate the use of reverse inferences from all brain regions which have been employed by cognitive scientists in recent studies. For example, the insula, which plays a crucial role in Case Study 2 and many other studies of higher-cognition, is especially problematic for reverse inferences (Chang et al. 2013). Our point is simply that each particular reverse argument must be evaluated in this task-relative way. Why is this requirement systematically overlooked? Part of the reason seems to be that in methodological discussions—such as Poldrack's discussion of Broca's area and language function and Phelps' discussion of the amygdala and negative emotions—theorists only consider arbitrary 'empty' tasks which do not eliminate any processing possibilities (i.e., any bridge-laws) for the brain region of interest. As a result, they get intuitively weak particular reverse inferences. What we are arguing is that if we seriously consider the set of particular tasks relevant to each reverse argument, we will usually have to eliminate some subset of the bridge-laws which cover the brain regions of interest, thereby increasing the strength of particular reverse inferences.[12]

An anonymous reviewer suggests that this task-relative account of reverse arguments faces a problem, namely, that it seems to assume that advocates of the competing hypotheses will agree on the subset of tasks over which to restrict the prior probability of activation of the regions of interest. But what if they disagree about the subset of tasks which activate the regions of interest? This

[12]There is one simple point worth mentioning because it is sometimes overlooked, giving rise to mistaken objections against particular reverse arguments. The success of a reverse argument depends, we have seen, on the selectivity of its brain regions relative to the relevant tasks. But it also depends on the degree of confirmation that each of the subcomponents established in Premise 3 give to each of the competing mechanisms (e.g., M and M^*). Assume for simplicity that the reverse argument only established the engagement of one subcomponent, m_1. The relevant values are determined by the following conditional probabilities:

$$P(M|m_1) = \frac{P(m_1|M)P(M)}{P(m_1)} \tag{2}$$

$$P(M^*|m_1) = \frac{P(m_1|M^*)P(M^*)}{P(m_1)} \tag{3}$$

(2) and (3) determine the conditional probability of engagement of each of the competing mechanisms, M and M^*, given engagement of the cognitive subcomponent m_1. Recall that in the first case study we said that the engagement of negative emotions such as fear in *footbridge* favors M_E over M_R, despite the fact that M_R is *compatible*, but does not require, the engagement of negative emotions. This is what (2) and (3) make clear, i.e., why the engagement of m_1 will always increase the confidence for the engagement of M, of which it is a necessary subcomponent, more than for M^*, of which it is not a subcomponent but is merely compatible. Assume $P(M)$=$P(M^*)$=0.5, i.e., that there is no reason to hold that one of the competing mechanisms is more plausible than the other (remember we are, somewhat artificially, examining reverse arguments in isolation). It is easy to see that $P(m_1|M) > P(m_1|M^*)$: m_1 is a subcomponent of M, so $P(m_1|M) = 1$, and m_1 is not a subcomponent of M^*, so $P(m_1|M^*) < 1$. Finally, note that often theorists care not only about which of the competing hypotheses is favored by the evidence, but also about the extent to which our confidence in the hypotheses is increased by the evidence. To determine whether this increase is substantial, we can use Bayes factor (Poldrack 2006), which is given by the ratio of the posterior odds to the prior odds, where the odds are determined by $p/(1-p)$.

legitimate worry concerns how to determine the set of bridge-laws which apply to the relevant brain regions, e.g. the set of laws that associate the amygdala with different cognitive functions. We agree that some disagreement about which bridge-laws cover a given brain-region is inevitable. For this reason, when a study is explicitly designed to compare competing hypotheses, experimenters have to be particularly careful to take into account a wide range of (plausible) bridge-laws for each brain region of interest. The key then is to devise a set of tasks, including control tasks, that allows them to test which bridge-laws apply in the relevant reverse argument.[13]

4. To say that reverse inferences should be evaluated relative to particular tasks is not necessarily to say that they should be further conditionalized on those tasks. Reformulating equation (1) by conditionalizing on tasks is one way of formally implementing the task-relativity of reverse inferences, but there are other and arguably more plausible options.

Recently, Hutzler (2013) proposed one way of formalizing similar observations about the task-relativity of reverse inferences. We fully agree with the rationale that motivates Hutzler's proposed revision, but have some doubts about his specific proposal. According to Hutzler, we should revise (1) by explicitly conditionalizing on the relevant task, say t_1:

$$P(m_1|n_1\&t_1) = \frac{P(n_1|m_1\&t_1)P(m_1|t_1)}{P(n_1|m_1\&t_1)P(m_1|t_1) + P(n_1|\neg m_1\&t_1)P(\neg m_1|t_1)} \qquad (4)$$

To intuitively motivate this revision, Hutzler presents a simple thought experiment (here reformulated in our terminology). Imagine that activation in the left fusiform gyrus ($= n_1$) is covered by two bridge-laws: 'Br_1' associates n_1 with access to the mental lexicon and 'Br_2' associates n_1 with face perception. Assume that there is a visual word presentation task t_1 that results in n_1. The question is whether n_1 significantly increases one's confidence that t_1 engages a processes of accessing the mental lexicon. If we use equation (1), then the increase in confidence in the hypothesis is diminished by the existence of Br_2, according to which n_1 can also signal face perception processes. But this is counterintuitive, for t_1 clearly has nothing to do with face perception processes. In contrast, if

[13] To illustrate, consider the (in)famous study of swing voters' reactions to videos of US presidential candidates (Iacoboni et al. 2007). What is especially objectionable about this study is precisely that the experimenters ignored various relevant bridge-laws, namely, bridge-laws that connected to obvious competing explanations of the data. For example, the experimenters interpreted increased activation in the amygdala when viewing Mitt Romney compared to when viewing the other candidates as suggesting that Romney produced in such subjects a relative increase in anxiety, fear, or some comparable negative emotion. But given our current knowledge of the bridge-laws that cover the amygdala, such differential activation could just as plausibly mean that Romney was, at the time, a less-known or more attractive candidate than the others. Had Iacoboni and colleagues taken these bridge-laws into account, they could have devised the appropriate control tasks (e.g. including videos of handsome-famous individuals, videos of handsome-unknown individuals, etc.) to discriminate between the various competing hypotheses.

we use equation (4), Br_2 becomes irrelevant: by taking t_1 into account we can eliminate the possibility that *in this case* n_1 underlies face perception.

Hutzler's proposal captures the intuitively correct result in this simple example, but it faces some difficulties in more realistic reverse inferences. Consider again Case Study 1, focusing for simplicity just on the amygdala. Note that, intuitively, the inference that amygdala activation signals the engagement of negative emotions involves two types of tasks, each matched with a particular pattern of amygdala activation. The relevant evidence is that, relative to control scenarios, the amygdala was activated in *footbridge* but not in *switch*-like tasks. This differential task-activation pattern is what, in our informal discussion, allows us to eliminate various bridge-laws, and increase our confidence that processes involving negative emotions are engaged (call this hypotheses m_1). Assume that $+amygdala = n_1$ (so that $\neg n_1$ just means no differential activation in the amygdala); and that t_1 = *footbridge*-type tasks and t_2 = *switch*-type tasks. We cannot just conditionalize on this additional data in the manner proposed in (4), since we would then conditionalize on incompatible evidence, namely, two tasks and two *different* patterns of amygdala activation. This illustrates the underlying problem with (4): it represents tasks and locations of neural activation (or lack of) as if they were independent evidence of the same type.

For this reason, we think that a better implementation is to represent the evidence for reverse inferences as pairs of tasks and regions of potential activation, $< t_i, n_i >$. We can then properly model the evidence used in realistic reverse inferences. To illustrate, let $< t_1, n_1 >$ = increased neural activation in *footbridge*-type tasks; $< t_2, \neg n_1 >$ = no increased neural activation in *switch*-type tasks. The relevant value to determine then is $P(m_1 | < t_1, n_1 > \& < t_2, \neg n_1 >)$. We are now conditionalizing on compatible evidence, and not treating tasks and locations of potential activation as if they were independent evidence of the same type. In addition, this also clarifies why many bridge-laws are eliminated in good reverse arguments. For example, if we assume that *footbridge*-tasks are not more novel for most subjects than *switch*-like tasks (see discussion in footnote 11 above), then the bridge-law which maps amygdala activation to process related to novel stimuli is eliminated by the evidence, namely, $< t_1, n_1 > \& < t_2, \neg n_1 >$. This decreases the value of $P(< t_1, n_1 > \& < t_2, \neg n_1 > | \neg m_1)$—the false alarm rate—which in turn increases the value of $P(m_1 | < t_1, n_1 > \& < t_2, \neg n_1 >)$.

4 Conclusions

In this article, we elucidated and defended one of the main ways in which neuroimaging can bear on theories of higher-cognition, namely *via* reverse arguments. We conclude by presenting two implications of our account for issues at the psychology-neuroscience interface.

First, it is often assumed—more or less explicitly—that it is only when a 'higher-level' theory is at an advanced stage that it is useful to consider its relation to lower-level theories (Coltheart 2004; Harley 2004). In particular, since most mechanisms of higher-cognition are (still) only roughly understood,

this allegedly suggests that that the study of higher-cognition is insulated from neuroscience. One of the surprising implications of our analysis is that the conditions for effective reverse-arguments are often satisfied best at early stages of theorizing. This is because competing psychological theories at advanced stages often involve cognitive mechanisms that are implemented in overlapping neural locations, making it harder to distinguish them via neuroimaging techniques. In contrast, competing psychological theories at the early stages usually posit cognitive mechanisms whose neural bases are easier to discriminate—e.g. deontological rule application vs. negative emotional reaction, or positively biased misperception vs. negative emotional reaction.

Second, as noted above, researchers often hold that those who believes that neuroimaging bears directly on psychological theories is thereby committed to some substantial *reductionist* view of the relation between psychology and neuroscience. Again, our analysis of reverse arguments shows how a careful employment of neuroimaging techniques is perfectly compatible with the autonomy of psychology, for reverse inferences require only associationist—not reductionist—bridge-laws. Confusion about this point stems from a simplistic interpretation of Marr-levels. It is often assumed that questions about Marr-level 2 mechanisms are independent from issues of neural implementation, and that neuroscientific data is only relevant when neural implementation can be seriously considered. However, as our case studies illustrate, neuroimaging can be used to select amongst competing cognitive mechanisms independently of more advanced issues about their neural implementation.

To conclude, we should emphasize that while this essay examines one popular form of argument—namely, (location based) *reverse arguments*—there are other important ways in which neuroimagining can advance the study of higher-cognition. The other main type of argument, which is often ignored by critics but is becoming increasingly influential in cognitive neuropsychology (Miller 2008; Poldrack 2011; Kriegeskorte 2011), is based on inferring the engagement of psychological states or processes from specific *patterns* (rather than *locations*) of neural activation (Henson 2005, 2006). Location-based reverse inferences and pattern-based inferences have different uses and misuses for psychology. We believe they each deserve a detailed and independent discussion.

References

Aminoff, E., et al. (2009). The landscape of cognitive neuroscience: Challenges, rewards, and new perspectives. In M. Gazzaniga (Ed.), *The Cognitive Neurosciences* (4th ed.) (pp. 1255-1262). Cambridge, MA: MIT Press.

Ariely, D., Huber, J. & Wetenbroch, K. (2005). When do losses loom larger than gains? *Journal of Marketing Research*, *42*, 134–138.

Baron-Cohen, S. (2011). *The Science of Evil*. New York: Basic Books.

Brown, T. C. (2005). Loss aversion without the endowment effect, and other

explanations for the wta–wtp disparity. *Journal of Economic Behavior and Organization*, *57*(3), 367–379.

Caramer, C. (2005). Three cheers–psychological, theoretical, empirical–for loss aversion. *Journal of Marketing Research*, *42*, 129–133.

Chang, L. J., Yarkoni, T., Khaw, M. W. & Safey, A. G. (2013). Decoding the role of the insula in human cognition: Functional parcellation and large-scale reverse inference. *Cerebral Cortex*, *23*, 739–749.

Ciaramelli, E., Muccioli, M., Ladavas, E. & Di Pellegrino, G. (2007). Selective deficit in personal moral judgment following damaga to ventromedial prefrontal cortex. *Social Cognitive and Affective Neuroscience*, *2*, 84–92.

Coltheart, M. (2004). Brain imaging, connectionism, and cognitive neuropsychology. *Cognitive Neuropsychology*, *2*, 21–24.

Eisenberger, N. I., Lieberman, M. D., & Williams, K.D. (2003). Does rejection hurt? An fmri study of social exclusion. *Science*, *302*(5643), 290–292.

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

Gallistel, C. R. (2009). The neural mechanisms that underlie decision making. In P. W. Glimcher, C. F. Camerer, E. Fehr & R. A. Poldrack (Eds.), *Neuroeconomics: Decision Theory and the Brain* (pp. 419–424). London: Academic Press.

Glimcher, P. W., Camerer, C. F., Fehr, E., & Poldrack, R.A. (2009). Introduction: A brief history of neuroeconomics. In P. W. Glimcher, C. F. Camerer, E. Fehr & R. A. Poldrack (Eds.), *Neuroeconomics: Decision Theory and the Brain* (pp. 1–12). London: Academic Press.

Greene, J. (2008). *The secret joke of Kant's soul*. In W. Sinnot-Armstrong, *Moral Psychology, Volume 3* (pp. 35–81). Cambridge, MA: MIT Press.

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

Greene, J. D., Sommerville R., Nystrom, L., Darley, J. & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*, 2105–2108.

Greene, J. D., Nystrom, L., Engell, A., Darley, J. & Cohen,J.D. (2004a). The neuronal bases of cognitive conflict and control in moral judgment. *Neuron*, *44*, 389–400.

Harley, T. (2004). Promises, promises. *Cognitive Neuropsychology*, *21*, 51–56.

Hauser, M. D. (2006). *Moral Minds*. New York: Harper Collins.

Henson, R. (2005). What can functional neuroimaging tell the experimental psychologist? *Quarterly Journal of Experimental Psychology*, *58A*, 193–233.

Henson, R. (2006). Forward inference using functional neuroimaging: Dissociations versus associations. *Trends in Cognitive Sciences*, *10*(2), 64–69.

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

Iacoboni, M., Freedman, J., Kaplan, J., Jamieson, K., Knapp, T. & Fitzgerald, K. (2007, November 11). This is your brain on politics. *The New York Times*.

Jeffreys, H. (1961). *Theory of Probability*. Oxford: Clarendon Press.

Joyce, R. (2008). What neuroscience can (and cannot) contribute to metaethics. In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Volume 3* (pp. 371–394). Cambridge: MIT Press.

Kahneman, D., Knetch, J. L. & Thaler R. H.(1990a). Anomalies: The endowment effect, loss aversion and status quo bias. *The Journal of Economic Perspective*, *5*(1), 193–206.

Kahneman, D., Knetch, J. L. & Thaler R. H.(1990b). Experimental tests of the endowment effect and the coarse theorem. *Journal of Political Economy*, *98*(9), 1325–1348.

Kahneman, D. & Novemsky, N. (2005). The boundaries of loss aversion. *Journal of Marketing Research*, *42*, 119–128.

Knutson, B. & Greer S. M. (2008). Anticipatory affect: Neural correlates and consequences for choice. *Philosophical Transactions of the Royal Society of London, B*, *363*(1511), 3771-3786.

Knutson, B., Wimmer, E. G., Rick, S., Hollon, N. G., Prelec, D., & Loewenstein, G. (2008). Neural antecedents and the endowment effect. *Neuron*, *58*, 814–822.

Koenings, M., Young, L., Adolphs, R., Tranel, D., Cushman, F. & Hauser, M. (2007). Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature*, *446*(908-911).

Kohlberg, L. (1971). From 'is' to 'ought': How to commit the naturalistic fallacy and get away with it in the study of moral development. In T. Mischel (Ed.), *Cognitive development and epistemology* (pp. 151–284). New York: Academic Press.

Kriegeskorte, N. (2011). Pattern-information analysis: From stimulus decoding to computational-model testing. *Neuroimage*, *56*, 411–421.

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

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E. & Feldman Barret, L. (2012). The brain basis of emotion: a meta-analytic review. *Behavioral and Brain Sciences*, *35*, 121–202.

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

Masten, C. L., Eisenberger, N. I., Borofsky, L. A., Pfeifer, J. H., McNealy, K. & Mazziotta J. C. (2009). Neural correlates of social exclusion during adolescence: Understanding the distress of peer rejection. *Social Cognitive and Affective Neuroscience*, *4*(2), 143–157.

Mendez, M. F., Anderson, E. & Shapira,J. S. (2005). An investigation of moral judgment in frontotemporal dementia. *Cognitive and Behavioral Neurology*, *18*, 193–197.

Mikhail, J. (2008). *Moral cognition and computational theory* In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Volume 3* (pp. 81-93). Cambridge: MIT Press.

Miller, E. K. & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202.

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

Moll, J., Olivera-Souza, R., Zahn, R. & Grafman,J. (2008). The cognitive neuroscience of moral emotions. In W. Sinnott-Armstrong (Ed.), *Moral Psychology*, Volume 3 (pp. 1–18). Cambridge, MA: MIT Press.

Morewedge, C. K., Shu, L. L., Gilbert, T. D. & Wilson, D. T. (2009). Bad riddance or good rubbish? Ownership and not loss aversion causes the endowment effect. *Journal of Experimental Social Psychology*, *45*(4), 947–951.

Nayakankuppam, D. & Mishra, H. (2005). The endowment effect: Rose-tinted and dark-tinted glasses. *Journal of Consumer Research*, *32*, 390–395.

Novemsky, N. & Kahneman, D. (2005). The boundaries of loss aversion. *Journal of Marketing Research*, *42*, 119–128.

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* (pp. 233–250). London: Academic Press.

Phelps, E. & Delgado, M. (2009). Emotion and decision making. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences* (4th ed.) (pp. 1093–1105). Cambridge, MA: MIT Press.

Poeppel, D. (1996). A critical review of PET studies of phonological processing. *Brain and Language*, *55*, 317–351.

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

Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: from reverse inference to large scale decoding. *Neuron*, *72*, 692–697.

Poldrack, R. A. & Wagner, A. (2004). What can neuroimaging tell us about the mind? Insights from prefrontal cortex. *Current Directions in Psychological Science*, *13*, 177–181.

Rick, S. (2011). Losses, gains, and brains: Neuroeconomics can help to answer open questions about loss aversion. *Journal of Consumer Psychology*, *21*, 453–463.

Ross, D. (2008). Two styles of neuroeconomics. *Economics and Philosophy*, *24*, 473–483.

Sanfey, A., Rilling, J., Aronson,J., Nystrom, L. & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, *300*, 1755–1758.

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

Schoenbaum, G. & Roesch, M. (2005). Orbitrofrontal cortex, associative learning, and expectancies. *Neuron*, *47*, 633–636.

Sinnott-Armstrong, W. (Ed.) (2008a). *Moral Psychology. The Cognitive Science of Morality: Intuition and Diversity*, Volume 2. Cambridge, MA: Bradford, MIT Press.

Sinnott-Armstrong, W. (Ed.) (2008c). *Moral Psychology. The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, Volume 3. Cambridge, MA: Bradford, MIT Press.

Smith, E. E. & Jonides, J. (1997). Working memory a view from neuroimaging. *Cognitive Psychology*, *33*, 5–42.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, *59*, 204–217.

Trssoldi, P. E., Sella, F., Coltheart, M. & Umilta, C. (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.

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

Uttal, W. R. (2002). Précis of the new phrenology: The limits of localizing cognitive processes in the brain. *Brain and Mind*, *3*(2), 221–228.

Uttal, W. R. (2011). *Mind and Brain: A Critical Appraisal of Cognitive Neuropsychology*. Cambridge, MA: MIT Press.

Valdesolo, P. & DeSteno, D. (2005). Manipulations of emotional context shape moral judgment. *Psychological Science*, *17*(476-477).

Van Orden, G. C. & Paap, K. R. (1997). Functional neuroimages fail to discover pieces of mind in the parts of the brain. *Philosophy of Science*, *64*, S85–94.