frontiers in
PSYCHOLOGY

REVIEW ARTICLE
published: 27 July 2011
doi: 10.3389/fpsyg.2011.00162

# Architecture of explanatory inference in the human prefrontal cortex

***Aron K. Barbey[1,2,3]* and Richard Patterson[4]***

[1] *Decision Neuroscience Laboratory, University of Illinois at Urbana-Champaign, Champaign, IL, USA*
[2] *Center on Health, Aging, and Disability, University of Illinois at Urbana-Champaign, Champaign, IL, USA*
[3] *Department of Speech and Hearing Science, University of Illinois at Urbana-Champaign, Champaign, IL, USA*
[4] *Department of Philosophy, Emory University, Atlanta, GA, USA*

***Edited by:***
*Diane Pecher, Erasmus University Rotterdam, Netherlands*

***Reviewed by:***
*Diane Pecher, Erasmus University Rotterdam, Netherlands*
*Chad Edward Forbes, University of Delaware, USA*

****Correspondence:***
*Aron K. Barbey, Decision Neuroscience Laboratory, 110 Huff Hall, 1206 South Fourth Street, Champaign, IL 61820 MC-586, USA.*
*e-mail: barbey@Illinois.edu*
*http://www.DecisionNeuroscienceLab.org/*

Causal reasoning is a ubiquitous feature of human cognition. We continuously seek to understand, at least implicitly and often explicitly, the causal scenarios in which we live, so that we may anticipate what will come next, plan a potential response and envision its outcome, decide among possible courses of action in light of their probable outcomes, make midstream adjustments in our goal-related activities as our situation changes, and so on. A considerable body of research shows that the lateral prefrontal cortex (PFC) is crucial for causal reasoning, but also that there are significant differences in the manner in which ventrolateral PFC, dorsolateral PFC, and anterolateral PFC support causal reasoning. We propose, on the basis of research on the evolution, architecture, and functional organization of the lateral PFC, a general framework for understanding its roles in the many and varied sorts of causal reasoning carried out by human beings. Specifically, the ventrolateral PFC supports the generation of basic causal explanations and inferences; dorsolateral PFC supports the evaluation of these scenarios in light of some given normative standard (e.g., of plausibility or correctness in light of real or imagined causal interventions); and anterolateral PFC supports explanation and inference at an even higher level of complexity, coordinating the processes of generation and evaluation with further cognitive processes, and especially with computations of hedonic value and emotional implications of possible behavioral scenarios – considerations that are often critical both for understanding situations causally and for deciding about our own courses of action.

**Keywords: explanatory inference, causal reasoning, executive control, lateral prefrontal cortex**

## PREFRONTAL CONTRIBUTIONS TO EXPLANATORY INFERENCE

The human mind is driven toward understanding. We wonder why events unfold in particular ways, why objects have specific properties and why people behave the way they do. The capacity to infer the causal structure of experience and to generate explanations is central to our sense of understanding, making possible the formation of conceptual representations that constrain inference, guide generalization, and provide the basis for goal-directed, intelligent behavior. Accordingly, extensive research in social psychology and philosophy is dedicated to the study of explanation, with social psychology focusing on explanations of behavior (e.g., Heider, 1958; Gilbert, 1998; Malle, 2004) and philosophy on explanation in science (e.g., Salmon, 1998; Kitcher and Salmon, 1989). Only recently, however, has cognitive science addressed such questions as what constitutes an explanation, what makes some explanations better than others, and the principles that determine when we seek explanations and how we generate them (e.g., Keil and Wilson, 2000; Keil, 2006).

Two recent developments have spurred the emergence and growth of research on explanation within cognitive psychology. First, prominent theories of conceptual knowledge accord a central role to explanation (e.g., Carey, 1985; Murphy and Medin, 1985; Murphy, 2002; Keil, 2003, 2006). Explanations facilitate category learning, influence judgments of the typicality of category members and foster conceptual coherence. Second, explanations and the causal representations they entail exert a profound influence on learning and reasoning (e.g., Barbey and Wolff, 2006, 2007, under review; Tenenbaum et al., 2006; Sloman et al., 2009; Wolff et al., 2010). Cognitive psychologists have therefore increasingly recognized the importance of investigating the psychology of explanation. We suggest that category learning, typicality judgments, reasoning, and conceptual coherence are strongly interconnected, and that our beliefs about the causal powers of objects, events, and agents – and about the rule-like causal relationships among them – are central to the generation and evaluation of the myriad ways in which we interpret, understand and explain ourselves and our environment.

Parallel developments in cognitive neuroscience have fostered the study of the neural mechanisms underlying explanation. For instance, the resurgence of cognitive simulation theories has motivated neuroscience models of explanatory inference based on the simulation of modality-specific components of experience (e.g., Damasio, 1989; Barsalou, 1999; Barbey and Barsalou, 2009; Barbey and Grafman, in press, 2011; Barbey et al., 2011a,b). According to this framework, a highly integrative, multimodal representation system in the brain supports simulation mechanisms for explanatory inference across the spectrum of cognitive activities, including high-level perception, implicit memory, working memory, long-term memory, and conceptual knowledge (for a recent review, see Barbey and Barsalou, 2009). Although the functional topography

of modality-specific representations and their role in these processes have become increasingly well understood, remarkably little is known about the simulation mechanisms that encode the higher-level structure of experience, representing causal relationships that support explanatory inference and establish the proper mappings between situations, actions and consequences necessary for coordinated, purposeful behavior. The absence of such data represents a substantial gap in understanding both the neural architecture of cognitive simulations and their role in higher cognitive functions.

Here we introduce an integrative cognitive neuroscience theory for understanding the mechanisms that enable the top-down control and coordination of modality-specific representations, drawing upon recent theoretical developments in cognitive psychology and emerging neuroscience evidence indicating that the lateral prefrontal cortex (PFC) supports the generation, evaluation, and coordination of representations that encode the causal structure and probable causal implications of events, and thus provide the basis for causal understanding of our environment and for our day-to-day navigation through that environment. We regard explanation featuring a specific causal mechanism as the central or prototypical case of causal explanation. Less central are cases in which one must act without having settled confidently on one mechanism or another, or again, while simply assuming that there is some causal mechanism at work but without having any good idea what it might be. At the ragged edge of understanding we sometimes have to be content with statistical correlations. Here we note another significant gap in current understanding of causal explanations: statistical correlations do in many circumstances give rise to the formulation of causal generalizations (including causal mechanisms) that we then apply to new cases in relevantly similar circumstances. And although we will review evidence that different brain systems support inductive reasoning to probabilistic generalizations on the one hand, and deductive causal reasoning (from a "major premise" asserting a general causal relationship and another premise bringing some particular event under that causal rule) on the other, the circumstances under which, and the processes by which, we move from statistical correlation to causality remain to be investigated.

We begin by reviewing psychological research on the structure of explanations, surveying contemporary research and theory from cognitive psychology suggesting that explanatory inference accommodates novel information in the context of background beliefs, as it enables generalizations and predictions about self, others, and the environment. We then review the biology, evolution and ontogeny of the human PFC, and introduce a cognitive neuroscience framework for causal inference based on a functional division of labor within the lateral PFC. Our review examines a broad range of evidence from the social and decision neuroscience literatures demonstrating that computational mechanisms for the generation and evaluation of causal simulations are mediated by functionally specialized regions of the lateral PFC (ventrolateral PFC and dorsolateral PFC, respectively), and that at yet higher levels of complexity, where these and other cognitive processes must be coordinated, causal inference is further supported by the anterolateral PFC. We show how this framework supports the integration of a diverse body of neuroscience evidence concerning human reasoning not just about basic physical and social contexts, but also within the context of moral, ethical, and legal systems of value and belief.

## PSYCHOLOGY OF EXPLANATION

Psychological evidence supports the predominance of causation in explanation (e.g., Barbey and Wolff, 2006, 2007, under review; Tenenbaum et al., 2006; Sloman et al., 2009; Wolff et al., 2010). Explanations typically appeal to causes, along with knowledge of general patterns that constrain which causes are judged to be probable (Einhorn and Hogarth, 1986) and relevant (Lombrozo and Carey, 2006; Wellman and Liu, 2007). Explanations recruit a great deal of prior knowledge, establishing relevant causal mechanisms that provide a source of constraint for reasoning and a basis for generalizing from known to novel cases (reviewed in Lombrozo, 2006). As a consequence, the top-down control and coordination of behavior depends on the capacity to generate causal explanations and understanding of the physical and social world.

## CAUSAL REPRESENTATIONS

A major function of the PFC is to extract statistical regularities across experience in an effort to infer general patterns and causal relationships that establish the proper mappings between situations, actions and consequences necessary for goal-directed behavior (for reviews, see Miller, 2000; Miller and Cohen, 2001; see also Barbey et al., in press). By extracting the higher-level structure of experience, representations within the PFC enable the top-down control and coordination of multiple brain mechanisms across diverse brain areas and networks (for a recent review of the neurobiological mechanisms underlying PFC function, see Miller and Phelps, 2010).

## CAUSAL REPRESENTATIONS IN THE LATERAL PFC

Behaviorally relevant causal information about causal powers and causal associations and patterns, and the causal inferences these support, are encoded by diverse areas of lateral PFC (**Figure 1**; for reviews, see Miller, 2000; Miller and Cohen, 2001). We will not discuss here the issue of whether our representations of events, or of agents and objects and their causal powers, are through-and-through a matter of modal simulations (as when we see or imagine billiard balls colliding, levees breaking, etc.), or whether more schematic and abstract – and perhaps even amodal – representations are involved. Elsewhere we suggest and defend a pluralistic approach within which modal simulation is the evolutionarily oldest – and probably still the default – medium of causal reasoning, but on which one employs more or less schematic, and more or less abstract modes of reasoning depending on one's circumstances (Patterson and Barbey, in press). As will become evident, the framework we propose here will accommodate a wide variety of views about the ground-level nature of causal representations.

While the lateral PFC is a site of convergence for the synthesis of multimodal information from a wide range of brain systems (see below on connectivity of these regions), we propose that the neural architecture of the lateral PFC entails two pathways for explanatory inference. The ventrolateral PFC supports the generation and maintenance of causal simulations, relying upon computational mechanisms for detecting and encoding causal relationships. Within this framework, as a causal event is experienced repeatedly, its simulated components and the causal relationships linking them increase in potency. Thus when one component is perceived

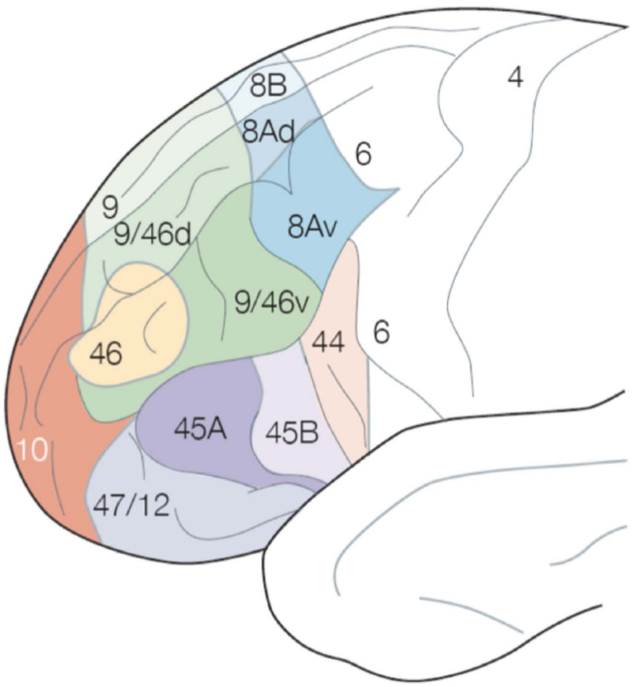


**FIGURE 1 | Brodmann map of the lateral PFC.** Reproduced with permission from Ramnani and Owen (2004).

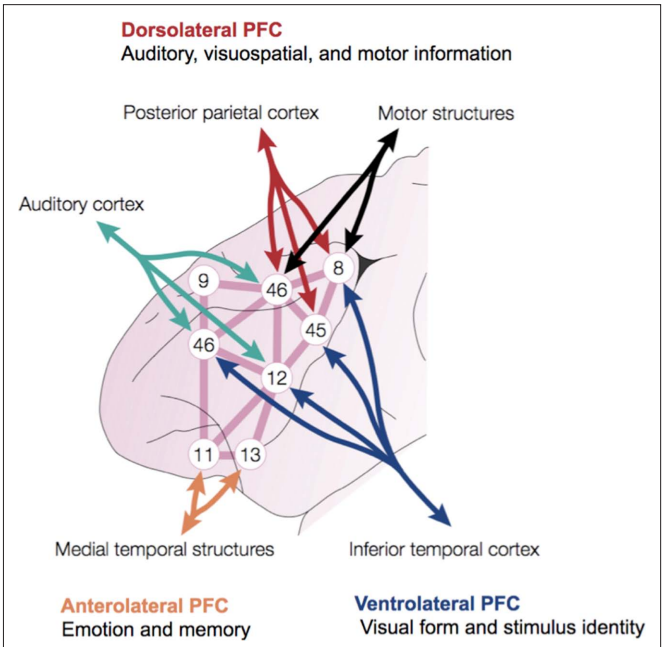


**FIGURE 2 | Integrative anatomy of the macaque monkey PFC.** Numbers refer to subregions within the lateral PFC defined by Brodmann. Modified with permission from Miller (2000).

initially, these strong associations complete the pattern automatically, supporting inferences about the underlying cause(s) and their resulting effect(s).

On the basis of this same experience one also forms representations of the causal powers, active and receptive, of the agents, objects, and events involved in causal situations. These representations underlie the implicit and explicit production of novel and counterfactual simulations critical for planning, monitoring, and adjusting behavior.

These first level explanations and inferences must often be evaluated and re-evaluated as we make our way in the world – e.g., by devising, imagining, or performing an intervention to find out whether or not an effect is present in the absence of the candidate cause. Such evaluations are supported by computational mechanisms in the dorsolateral PFC. This framework of causal evaluation operating over representations of causal patterns and powers is supported in the first instance by classic neuroscience research on working memory, which demonstrates that the ventrolateral PFC supports the maintenance of cognitive representations and the dorsolateral PFC is additionally recruited for monitoring and manipulating items (e.g., Petrides, 2005; D'Esposito et al., 1999). Further, and more direct, evidence is reviewed below.

## ANATOMICAL CONNECTIVITY, EVOLUTION, AND DEVELOPMENT OF THE LATERAL PFC

The inferential architecture of the lateral PFC derives from the anatomical connectivity, evolution, and development of this region. The lateral PFC consists of three major subregions that emphasize processing of particular information based on their interconnections with specific cortical regions (**Figure 2**).

Ventrolateral areas are more heavily interconnected with cortical regions for processing information about visual form and stimulus identity (inferior temporal cortex) that supports the detection of causal relationships and the categorization of environmental stimuli. Dorsal portions of the lateral PFC are heavily interconnected with cortical areas for processing visuospatial and motor as well as auditory information. It is primarily the capacity to manipulate visuospatially arrayed representations of objects and events in causal scenarios that makes possible the evaluation and adjustment of causal understanding to serve one's short and long-term needs (Barbey et al., 2009a). Finally, the anterolateral PFC is indirectly connected (via the ventromedial PFC) with limbic structures that process internal information, such as emotion, memory, and reward (Goldman-Rakic, 1987; Pandya and Barnes, 1987; Fuster, 1989; Barbas and Pandya, 1991). The lateral PFC therefore enables the synthesis of information across this broadly distributed network of modal brain regions (for pertinent reviews, see Kringelbach, 2005).

Research investigating the evolution and ontogeny of the PFC suggests that the lateral PFC initially emerged from ventrolateral prefrontal regions, followed by dorsolateral and then anterolateral cortices (**Figure 3**; Flechsig, 1901, 1920; Fuster, 1997).

From an evolutionary perspective, the emergence of lateral PFC subregions reflects their relative priority for goal-directed behavior, with the ventrolateral PFC enabling the capacity to maintain basic causal beliefs and generate explanatory inferences. The fine details of the human capacity to represent a range of possible causal outcomes or antecedents of a given situation are not yet understood, but we suggest that these go hand-in-hand with the capacity to evaluate causal explanations and to plan, monitor, and adjust causal behavior in light of our causal understandings – abilities supported by the dorsolateral PFC.

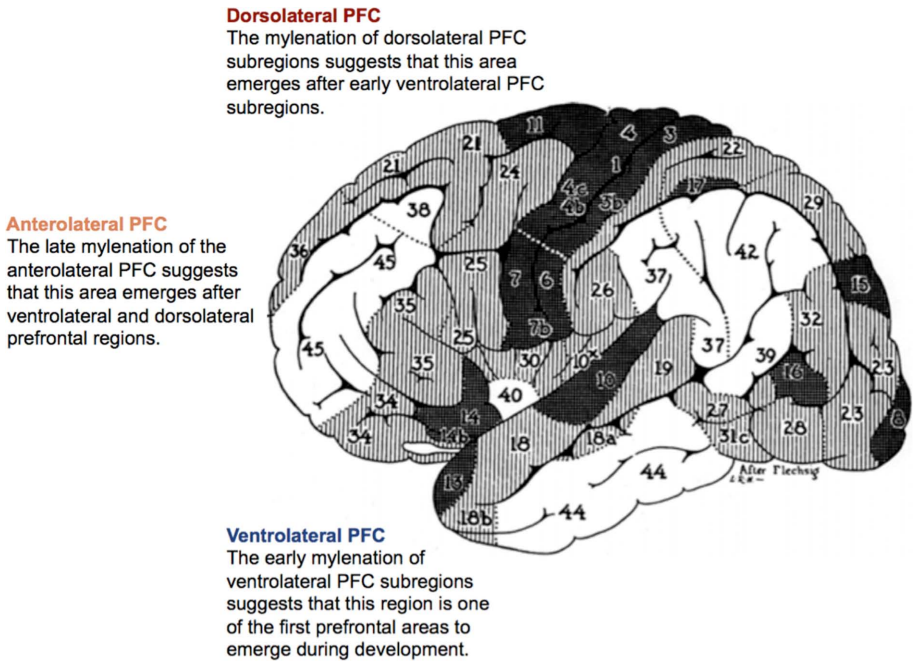


**FIGURE 3 | Ontogenetic map of the prefrontal cortex according to Flechsig (1901, 1920).** The numeration of the areas indicates the order of their myelination. Modified with permission from Flechsig (1920).

Finally, the evolution of the anterolateral PFC enabled processing of higher-order relations and reasoning about complex forms of goal-directed behavior involving both the generation and evaluation of explanatory inferences (for a review, see Ramnani and Owen, 2004), but also the integration of these processes with hedonic and emotional information associated with different causal scenarios, and especially with different possible causal outcomes. Consistent with its evolutionary development, the ontogeny of the lateral PFC reflects the importance of first establishing explanations for understanding the physical and social world, followed by the capacity to manipulate and evaluate these explanations, and finally high-order inferences involving both sorts of activity – along with the coordination of these processes with further relevant information and computation including the assessment of hedonic outcomes of possible actions. This coordination of multiple processes will routinely characterize human inferences about the multifaceted (causal) outcomes of actions, and is in general supported by the anterolateral PFC (Ramnani and Owen, 2004). We focus here on the anterolateral PFC, but with some reference to its connections with the larger anterior PFC and orbitofrontal cortex (OFC), among other regions. Hedonic valences of rewards and punishers along with their connections to specific stimuli are represented largely in OFC, and this information will typically be incorporated into human calculations of outcomes and decision making (Kringelbach, 2005). For a recent review of pertinent developmental evidence, see Rochat (2009).

This is not to suggest that any of such functions (e.g., stimulus identification via spatial properties, spatial mapping of the environment, control of behavior, engagement in social transactions, etc.) are carried out solely by any one region of PFC. On the contrary, physical, biological, emotional, social, and other information will be used in performing the functions we associate with all three areas of PFC focused on here. The suggestion is rather that these functions and their anatomical correlates are integrated, and that there is substantial evidence that one useful way of distinguishing functions of the PFC with regard to causal explanation and inference in particular coincides with the anatomical division into ventrolateral, dorsolateral, and anterolateral PFC.

An empirical case for the functional specificity of ventrolateral PFC for the maintenance of information and dorsolateral PFC in the manipulation of representations has been established in the functional neuroimaging literature on working memory, providing evidence that broadly supports the proposed functional organization of lateral PFC (for meta-analytic reviews, see Wager and Smith, 2003; Wager et al., 2004; Owen et al., 2005). It makes good sense that dorsolateral PFC is, on the one hand, heavily involved in the evaluation of causal explanations and, on the other hand, is strongly interconnected with visuospatial processing areas, because such evaluation is typically carried out via manipulation of visuospatial arrays of representations of potential causes, background conditions, enablers, etc. This holds for theories emphasizing the use of modal simulations as well as for those centering on much more schematic, abstract, and even amodal representations (see Barbey and Wolff, 2006, 2007; Patterson and Barbey, in press, under revision; Sloman et al., 2009). Meanwhile anterolateral PFC has been shown in a range of research reviewed above to be involved on the one hand in higher-order reasoning and, on the other, in social cognition. This correlation, too, makes good sense in that human social understanding and decision making are frequently complex, drawing on and integrating multiple cues of diverse types

into coherent explanatory scenarios. "Immediate" situations are in turn imbedded in larger causal scenarios and narratives that one must take into account, where these sometimes reach as far as overarching life goals, and where one wants to consider at many points the likely hedonic and emotional impact of possible actions. And of course, such representations are subject to both top-down and bottom-up influences, as we consider "what difference it might make" for our pursuit of some larger goal if we undertake one immediate action rather than another, or for what we should do here and now if we are to further one long-term goal rather than another. This further suggests that for the anterolateral PFC to fulfill its role not just with regard to higher-order inference in general, but social and emotional life in particular, its strong connectivity to OFC and, via ventromedial PFC, with the limbic system is critical, since our inferences about what will result from a given action, and for whom, and how, will have to include much information about, and computations of, hedonic and emotional values. But the evolutionary history of the connectivity of the anterolateral PFC with OFC and subcortical limbic areas remains to be written.

We note finally that the picture given just above dovetails with additional evidence that the anterior-to-posterior axis of the lateral PFC is organized hierarchically, whereby progressively anterior subregions are associated with higher-order processing requirements for planning and the selection of action (for recent reviews, see Ramnani and Owen, 2004; Koechlin and Summerfield, 2007; Badre, 2008; Botvinick, 2008). Thus, processes within the lateral PFC respect the hierarchical organization of this region, with progressively anterior regions representing causal simulations that support higher-order inferences based on computational mechanisms for generating and evaluating explanations.

## INFERENTIAL ARCHITECTURE OF THE LATERAL PFC

We now turn to a review of recent evidence from the social and decision neuroscience literatures demonstrating (1) the involvement of the ventrolateral PFC for the generation and maintenance of explanatory inferences, (2) the recruitment of the dorsolateral PFC for evaluating possible explanations in light of normative criteria, and (3) activation in the anterolateral PFC for manipulation and utilization of higher-order inferences that incorporate both types of process and also coordinate these with other relevant processes, such as computation of hedonic values of predicted outcomes of potential actions (**Figure 4**). The representational architecture underlying these forms of inference further predicts the recruitment of broadly distributed neural systems, incorporating medial prefrontal (Barbey et al., 2011a; for reviews, see Kringelbach, 2005; Amodio and Frith, 2006) and posterior knowledge networks (e.g., Simmons et al., 2010) representing unimodal and multimodal components of experience.

### VENTROLATERAL PFC

Social neuroscience studies have shown that explanatory inference is mediated by the ventrolateral PFC (areas 44, 45, and 47; **Figure 4B**). Fiddick et al. (2005), observed recruitment of ventrolateral PFC when participants drew explanatory social inferences based on normative beliefs concerning reciprocal altruism and social exchange. Speaking more generally, this region is recruited when representing normative rules that guide social behavior (Barbey et al., 2009a). It is particularly sensitive to norm violations that motivate explanatory inferences about the cause(s) of deviant (i.e., non-normative) behavior (for recent behavioral evidence, see Uttich and Lombrozo, 2010). Berthoz et al. (2002), for

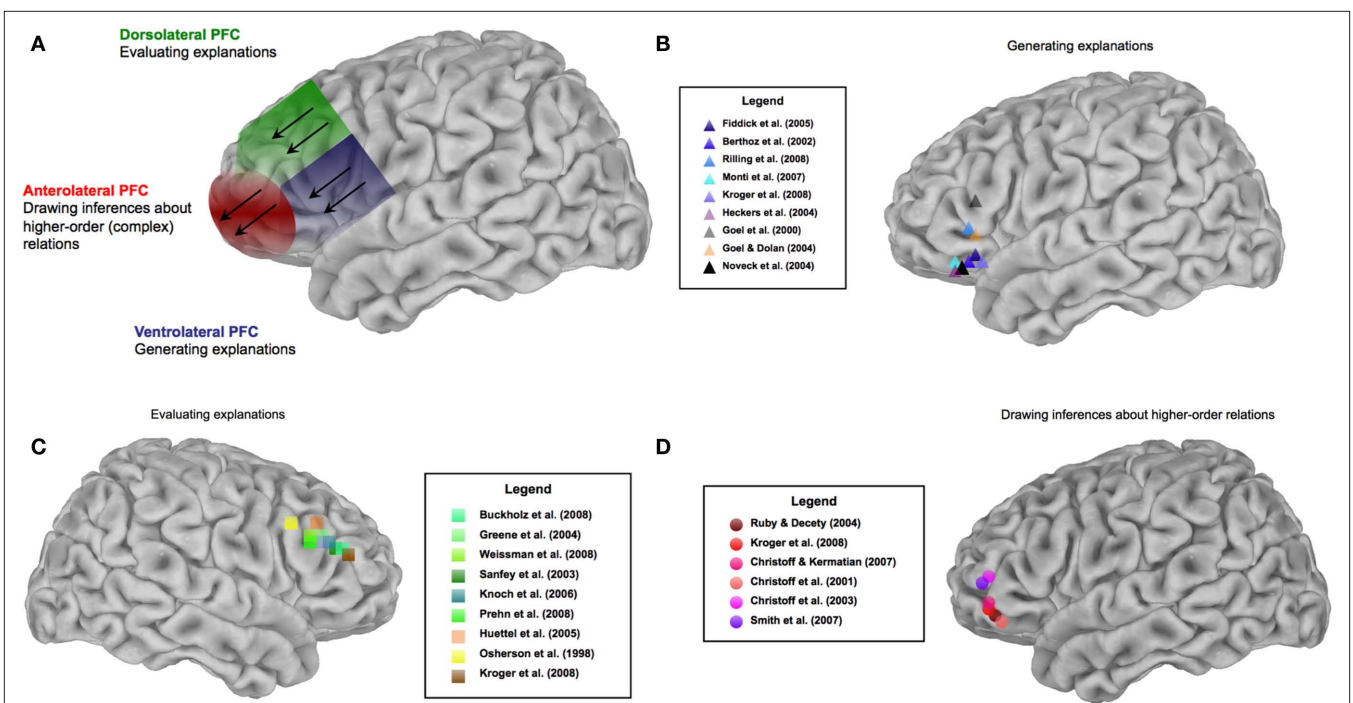


**FIGURE 4 | Neural architecture of explanatory inference. (A)** Summarizes the functional organization of the lateral PFC, and **(B–D)** illustrate supportive evidence.

example, demonstrated recruitment of left ventrolateral PFC (area 47) when participants detected violations of social norms stories representing obligatory and prohibited courses of action (e.g., the decision to "spit out food made by the host"). Similarly, Rilling et al. (2008) reported activation within left ventrolateral PFC (area 47) when participants detected the failure to cooperate in a Prisoner's dilemma game. Here the need for explanation is especially pressing, for we want to understand why the usual explanation or cause (as set forth in a social norm or rule) does not hold in a particular case. Does some "higher" or super-ordinate rule – or simply a conflicting coordinate rule – come into play in this particular situation? Or is there some other explanation altogether? In the food-spitting example, the explanation is presumably not a matter of a conflict among social rules, but a visceral response (perhaps to food that a particular person finds intensely nauseating). The research here does not establish whether ventrolateral PFC can "handle" the testing and comparison of alternative explanations (as opposed to the detection of a violation of some salient and usually explanatorily adequate social rule or norm). This needs to be investigated further, as part of making as clear as possible the conditions under which dorsolateral PFC, or even anterolateral PFC, must be recruited.

Meanwhile the decision neuroscience literature supports the proposed tripartite framework by suggesting a reason why, from a wider perspective, the ventrolateral PFC should be involved in the generation of basic social (and other) sorts of causal explanation. One very common type of explanation in everyday life and in scientific contexts depends essentially on deductive inference, as when some "covering law," or a behavioral rule or norm, combines with a statement of some particular facts to entail that some other fact must obtain, or that some specific action is obligatory, etc. For example, if books burn at Fahrenheit 451, and this book is heated to Fahrenheit 500, it will burn – that is, given an implicit ceterus paribus clause ("other things being equal"), or given the appropriate "enabling" or "background" conditions, such as that the book is not sopping wet, etc. Similarly, a great many everyday explanations in social and psychological, as well as physical, domains will explicitly or implicitly take the form of very simple deductions. For example, Why does one not spit out this bite of food, even if it tastes bad? Because one must not offend one's host. Phrased as a very simple deduction, we have: one must not do something that offends the host; spitting out the host's home-cooked food would offend the host; therefore one must not spit out the host's home-cooked food. This gives a general explanation for a more specific, but still general, rule. To get a prohibition against *this* person's spitting out *this* host's food on *this* occasion, we simply note that this person is a guest of that person, and that this food was home-cooked by the latter. This casts the explanation of why one doesn't spit out the host's food as a natural and simple deduction appealing to a general premise about what is impermissible. The explanation of why someone on a particular occasion did spit out the host's food would appeal to a different premise and deduction. (Perhaps, One involuntarily spits out food that is rotten and intensely nauseating; *this* person finds *this* food intensely nauseating.) Many everyday deductions are so intuitively obvious that they can be carried out automatically, but others will require conscious attention. The important point here is that a substantial body of neuroscientific research, if not a complete consensus in the field, strongly indicates that when common sources of variability are controlled (regarding the linguistic content, linguistic complexity, and deductive complexity of reasoning problems), simple deductions in general are supported by ventrolateral PFC. A recent series of experiments by Monti et al. (2007) controlled for these sources of variability and provided evidence that the left ventrolateral PFC (area 47) mediates representations of the logical structure of a deductive argument (e.g., If *P* or *Q*, then *Not-R*/*P*/Therefore, *Not-R*), supporting the generation of explanatory inferences within this region. Furthermore, a recent study by Kroger et al. (2008) controlled for the complexity and type of calculations that were performed and also observed activation within the left ventrolateral PFC (areas 44 and 45) for deductive reasoning (see also Heckers et al., 2004). Converging evidence is provided by Goel and colleagues (Goel et al., 2000; Goel and Dolan, 2004), who have consistently observed activation within the left ventrolateral PFC (areas 44 and 45) for deductive conclusions drawn from categorical syllogisms (e.g., All humans are mortal/Some animals are human/Therefore, some animals are mortal). Finally, Noveck et al. (2004) demonstrated recruitment of left ventrolateral PFC (area 47) for drawing deductive conclusions from conditional statements (e.g., If *P* then *Q*/*P*/Therefore, *Q*), consistent with the role of this region for generating explanatory inferences. In sum, this evidence indicates that generating a broad array of physical, social, and other explanations are supported by the ventrolateral PFC.

### DORSOLATERAL PFC

Social neuroscience evidence demonstrates that the dorsolateral PFC (areas 46 and 9) represents computational mechanisms for evaluating explanations and causal scenarios based on some normative criterion, where that may involve testing an attribution of correctness to a causal scenario (as in thinking about pertinent causal interventions), or about an attribution of fairness or permissibility – or the opposite – concerning causal outcomes of alternative possible actions (**Figure 4C**). An early study by Sanfey et al. (2003) reported activity within the right dorsolateral PFC (area 46) when participants were presented with an unfair offer. Knoch et al. (2006) further demonstrated that deactivating this region with repetitive transcranial magnetic stimulation reduced participants' ability to reject unfair offers in an ultimatum game. In these cases the making of an offer is in itself in accordance with the norms or rules of the game, but a norm of fairness has been violated – perhaps egregiously, as when one is offered only 1 dollar out of the 10 to be divided. One's response goes beyond detection of the unfairness, since one must then decide what to do about accepting or rejecting the offer, and this will typically involve weighing the outcomes of these options in light of multiple goals – the goal of maximizing one's money in the game, or the social goal of asserting and maintaining one's status (Rilling et al., 2008), or the goal of defending a commitment to fairness and to punishing unfairness, and perhaps other "higher" or coordinate goals.

Buckholtz et al. (2008) observed activity within the right dorsolateral PFC (area 46) when participants evaluated the causal role of specific factors, assigning responsibility for crimes and making judgments about appropriate (e.g., equitable or fair) forms of punishment in a legal decision making task. The work of Greene et al. (2004) further suggests that this region is involved

in normative evaluations involving conflicting moral goals. These authors employed moral scenarios similar to the trolley problem (Thomson, 1976) and assessed trials in which participants acted in the interest of greater aggregate welfare at the expense of personal moral standards. This contrast revealed reliable activation within the right dorsolateral PFC (area 46), suggesting that this region is critical for normative evaluations involving conflicting moral goals. (For additional evidence for the role of this region in such evaluative processes, see Prehn et al., 2008; Weissman et al., 2008).

We suggest that the relevant common factor in all these various cases is a "second order" reflection upon an initial or "first order" causal scenario (whose formulation was supported by ventrolateral PFC), where reflection on that scenario is motivated by some need to "think again" or "think twice," in order to decide whether to attribute or withhold attribution of some normative property – e.g., moral permissibility, fairness, social utility – to some given causal scenario(s). We suggest that the main reason studies with widely differing orientations have found involvement of dorsolateral PFC is that the kinds of norms involved show a corresponding variety.

### ANTERIOR PFC

Additional support for the general framework proposed here derives from the decision neuroscience literature, which demonstrates that progressively anterior subregions of the lateral PFC (area 10) are associated with higher-order processing requirements for thought and action (Koechlin and Summerfield, 2007; Badre, 2008; Botvinick, 2008). Ramnani and Owen (2004) reviewed contemporary research and theory investigating the cognitive functions of the anterior PFC, concluding that this region is central for integrating the outcomes of multiple cognitive operations, consistent with the predicted role of the anterior PFC for representing higher-order inferences that depend on the generation and evaluation of explanatory inferences (for representative findings, see Christoff et al., 2001, 2003; Christoff and Keramatian, 2007; Smith et al., 2007; Kroger et al., 2008; Barbey et al., 2011a).

This is to frame the issue once again in terms of levels of processing, with complexity increasing with anteriority. A large body of social neuroscience evidence supports this picture, although it may at first glance appear in some cases to invoke anterior PFC on the basic of content or subject matter (social and emotional) rather than level of complexity. It is well-established that anterior PFC (areas 10 and 11) – and the OFC more broadly – are central for explanatory social inference (**Figure 4D**). Studies of patients with lesions confined to the OFC have reported impairments in a wide range of social functions, including the regulation and control of social responses, the perception and integration of social cues, theory of mind and perspective taking (Rolls et al., 1994; Bechara et al., 2000; Ruby and Decety, 2004; LoPresti et al., 2008). Recent evidence from Stone et al. (2002) further demonstrates that patients with OFC damage show selective impairments in reasoning about normative social behavior and drawing explanatory social inferences. Bechara et al. (2000) observed profound deficits in the ability of OFC patients to represent and integrate social and emotional knowledge needed to generate mental state ascriptions and explanatory inferences about the causes of observed social behavior. Converging evidence is provided by LoPresti et al. (2008), who demonstrated that the left anterolateral PFC (area 11) mediates the representation and assessment of multiple social cue (i.e., emotional expression and personal identity), further suggesting that this region broadly supports the generation and evaluation of explanatory social inferences (for additional neuroscience evidence, see Moll et al., 2006).

It is, however, not merely the social nature of such inference that calls for involvement of anterior PFC, for as noted earlier, some social inferences (involving both adherence to and violations of social norms) are primarily supported by ventrolateral or dorsolateral PFC. What marks the involvement of anterior PFC, and explains the strong connection to social reasoning, is that typical "real life" social inference involves coordination and integration of multiple cognitive or computational tasks in the service of a larger goal. In particular, social reasoning that guides actual interactions with others will routinely integrate emotional and hedonic considerations into the evaluation of potential explanations of past actions and into deliberation about potential courses of future action. Thus anterior PFC draws on the resources of limbic areas (with which it is strongly connected via ventromedial PFC) and on hedonic representations in OFC (with which it is strongly interconnected; Kringelbach, 2005), as well as on the explanatory scenarios and the evaluative reflections and manipulations supported by ventrolateral and dorsolateral PFC.

## CONCLUSION

### INFERENTIAL ARCHITECTURE OF LATERAL PFC

We have reviewed converging lines of evidence to support the central role of the lateral PFC in explanatory inference, drawing upon recent theoretical developments in cognitive psychology and neuroscience bearing on the biology, evolution, ontogeny, and cognitive functions of this region. We have surveyed a broad range of evidence from social and decision neuroscience demonstrating that the lateral PFC mediates the generation and evaluation of explanations, with the ventrolateral PFC recruited when constructing explanatory inferences, engagement of the dorsolateral PFC for the evaluation of explanations, and the anterolateral PFC recruited when we utilize both of these processes – and additional ones with which they must be coordinated (such as the calculation of hedonic values of possible outcomes or actions; **Figure 4A**). The reviewed findings set the stage for new approaches to understanding the architecture of cognitive understanding suggesting that neural mechanisms within the lateral PFC detect and encode the higher-level structure of experience, representing causal relationships that guide the selection and control of modality-specific knowledge and provide the basis for explanatory inference.

Our findings raise further questions for future neuroscience research. One challenge is to address how neural mechanisms for generating and evaluating explanatory inferences are represented within dual process theories that distinguish between automatic versus controlled cognitive processes (e.g., Barbey and Sloman, 2007). Future research should further investigate the cognitive operations that are performed within the lateral PFC to support human inference. Does this region contain mechanisms that control the recruitment of representations stored in posterior cortices (e.g., Barbey et al., 2009a,b, 2011a,b, in press; Barbey and Grafman, in press, 2011), serve as an integrative hub for synthesizing modality-specific representations (e.g.,

Pessoa, 2008), or store unique forms of knowledge (e.g., Wood and Grafman, 2003)? Future research should also address the biological, developmental and evolutionary principles that account for the observed lateralization of processes for generating (left hemispheric) versus evaluating (right hemispheric) explanations (**Figure 4**). Research should further investigate the computational mechanisms underlying hierarchical cognitive representations with particular emphasis on the computational principles that enable the top-down control and coordination of modality-specific representations. Further research is needed also to define more precisely the functional and neural boundaries between ventrolateral, dorsolateral, and anterolateral PFC, especially given their intraPFC interconnections and their cooperative functioning in all but the most simple situations. The evidence surveyed here supports a broadly drawn tripartite framework for PFC involvement in causal explanation and understanding, but much of this evidence is itself indirect, and not designed precisely to investigate involvement of subregions of lateral PFC in explanatory reasoning. Finally, future research should investigate the larger role of the lateral PFC in the formation of human belief systems, investigating the neural mechanisms that integrate networks of causal knowledge to construct complex systems of value and belief, providing the foundations for explanatory inference and our sense of understanding.

## REFERENCES


Amodio, D., and Frith, C. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nat. Rev. Neurosci.* 7, 268–277.

Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. *Trends Cogn. Sci. (Regul. Ed.)* 12, 193–200.

Barbas, H., and Pandya, D. (1991). "Patterns of connections of the PFC in the rhesus monkey associated with cortical architecture," in *Frontal Lobe Function and Dysfunction*, eds H. S. Levin, H. M. Eisenberg, and A. L. Benton (New York: Oxford University Press), 35–58.

Barbey, A. K., and Barsalou, L. W. (2009). "Reasoning and problem solving: models," in *Encyclopedia of Neuroscience*, Vol. 8, ed. L. Squire (Oxford: Academic Press), 35–43.

Barbey, A. K., Krueger, F., and Grafman, J. (2009a). An evolutionarily adaptive neural architecture for social reasoning. *Trends Neurosci.* 32, 603–610.

Barbey, A. K., Krueger, F., and Grafman, J. (2009b). Structured event complexes in the prefrontal cortex support counterfactual representations for future planning. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 1291–1300.

Barbey, A. K., and Sloman, S. A. (2007). Base-rate respect: from ecological rationality to dual processes. *Behav. Brain Sci.* 30, 241–297.

Barbey, A. K., and Wolff, P. (2006). "Causal reasoning from forces," in *Proceedings of the 28 Annual Conference of the Cognitive Science Society* (Mahwah, NJ: Lawrence Erlbaum), 2439.

Barbey, A. K., and Wolff, P. (2007). "Learning causal structure from reasoning," in *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (Mahwah, NJ: Lawrence Erlbaum), 713–718.

Barbey, A. K., and Grafman, J. (in press). "The prefrontal cortex and goal-directed social behavior," in *The Handbook of Social Neuroscience*, eds J. Decety and J. Cacioppo (Oxford University Press).

Barbey, A. K., and Grafman, J. (2011). An integrative cognitive neuroscience theory for social reasoning and moral judgment. *Wiley Interdiscip. Rev. Cogn. Sci.* 2, 55–67.

Barbey, A. K., Koenigs, M., and Grafman, J. (2011a). Orbitofrontal contributions to human working memory. *Cereb. Cortex* 21, 789–795.

Barbey, A. K., Krueger, F., and Grafman, J. (2011b). "Architecture of counterfactual thought in the prefrontal cortex," in *Predictions in the Brain: Using our Past to Prepare for the Future*, ed. M. Bar (New York: Oxford University Press), 40–57.

Barbey, A. K., Solomon, J., Colom, R., Krueger, F., Forbes, C., and Grafman, J. (in press). An integrative architecture for general intelligence and executive function revealed by lesion mapping. *Brain*.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behav. Brain Sci.* 22, 577–660.

Bechara, A., Damasio, H., and Damasio, A. (2000). Emotion, decision making, and the orbitofrontal cortex. *Cereb. Cortex* 10, 1047–3211.

Berthoz, S., Armony, J. L., Blair, R. J. R., and Dolan, R. J. (2002). An fMRI study of intentional and unintentional (embarrassing) violations of social norms. *Brain* 125, 1696–1708.

Botvinick, M. M. (2008). Hierarchical models of behavior and prefrontal function. *Trends Cogn. Sci. (Regul. Ed.)* 12, 201–208.

Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., and Marois, R. (2008). The neural correlates of third-party punishment. *Neuron* 60, 930–940.

Carey, S. (1985). *Conceptual Change in Childhood*. Cambridge, MA: Bradford Books, MIT Press.

Christoff, K., and Keramatian, K. (2007). "Abstraction of mental representations: theoretical considerations and neuroscientific evidence," in *The Neuroscience of Rule-Guided Behavior*, eds S. A. Bunge and J. D. Wallis (New York: Oxford University Press), 107–128.

Christoff, K., Prabhakaran, V., Dorfman, J., Zhao, Z., Kroger, J. K., Holyoak, K. J., and Gabrieli, J. D. E. (2001). Rostrolateral PFC involvement in relational integration during reasoning. *Neuroimage* 14, 1136–1149.

Christoff, K., Ream, J. M., Geddes, L. P. T., and Gabrieli, J. D. E. (2003). Evaluating self-generated information: anterior prefrontal contributions to human cognition. *Behav. Neurosci.* 117, 1161–1168.

Damasio, A. R. (1989). Time-locked multiregional retroactivation: a systems-level proposal for the neural substrates of recall and recognition. *Cognition* 33,25–62.

D'Esposito, M. D., Postle, B. R., Ballard, D., and Lease, J. (1999). Maintenance versus manipulation of information held in working memory: an event-related fMRI study. *Brain Cogn.* 41, 66–86.

Einhorn, H. J., and Hogarth, R. M. (1986). Judging probable cause. *Psychol. Bull.* 99, 1–19.

Fiddick, L., Spampinato, M. V., and Grafman, J. (2005). Social contracts and precautions activate different neurological systems: an fMRI investigation of deontic reasoning. *Neuroimage* 28, 778–786.

Flechsig, P. (1901). Developmental (myelogenetic) localisation of the cerebral cortex in the human subject. *Lancet* 2, 1027–1029.

Flechsig, P. (1920). *Anatomie des Menschlichen Gehirnsund Ruckenmarks auf Myelogenetischer Grundlage*. Leipzig: Thieme; New York: Basic Books.

Fuster, J. M. (1989). *The PFC*. New York: Raven.

Fuster, J. M. (1997). *The PFC–Anatomy Physiology, and Neuropsychology of the Frontal Lobe*. Philadelphia: Lippincott-Raven.

Gilbert, D. T. (1998). "Ordinary personology," in *The Handbook of Social Psychology*, 4th Edn, eds D. T. Gilbert, S. T. Fiske, and G. Lindzey (New York: Oxford University Press), 89–150.

Goel, V., Buchel, C., Frith, C., and Dolan, R. (2000). Dissociation of mechanisms underlying syllogistic reasoning. *Neuroimage* 12, 504–514.

Goel, V., and Dolan, R. J. (2004). Differential involvement of left PFC in inductive and deductive reasoning. *Cognition* 93, B109–B121.

Goldman-Rakic, P. S. (1987). "Circuitry of primate PFC and regulation of behavior by representational memory," in *Handbook of Physiology: The Nervous System*, ed. F. Plum (Bethesda, MD: American Physiology Society), 373–417.

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

Heckers, S., Zalesak, M., Weiss, A. P., Ditman, T., and Titone, D. (2004). Hippocampal activation during transitive inference in humans. *Hippocampus* 14, 153–162.

Heider, F. (1958). *The Psychology of Interpersonal Relations*. New York: John Wiley & Sons, Inc.

Keil, F. C. (2003). Folkscience: coarse interpretations of a complex reality. *Trends Cogn. Sci. (Regul. Ed.)* 7, 368–373.

Keil, F. C. (2006). Explanation and understanding. *Annu. Rev. Psychol.* 57, 227–254.

Keil, F. C., and Wilson, R. A. (2000). *Explanation and Cognition*. Cambridge, MA: MIT Press.

Kitcher, P., and Salmon, W. (eds). (1989). *Scientific Explanation: Minnesota Studies in the Philosophy of Science*, Vol. 13. Minneapolis: University of Minnesota Press.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., and Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right PFC. *Science* 314, 829–832.

Koechlin, E., and Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends Cogn. Sci. (Regul. Ed.)* 11, 229–235.

Kringelbach, M. L. (2005). The human orbitofrontal cortex: linking reward

to hedonic experience. *Nat. Rev. Neurosci.* 6, 691–702.
Kroger, J. K., Nystrom, L. E., Cohen, J. D., and Johnson-Laird, P. N. (2008). Distinct neural substrates for deductive and mathematical processing. *Brain Res.* 1243, 86–103.
Lombrozo, T. (2006). The structure and function of explanations. *Trends Cogn. Sci. (Regul. Ed.)* 10, 464–470.
Lombrozo, T., and Carey, S. (2006). Functional explanation and the function of explanation. *Cognition* 99, 167–204.
LoPresti, M. L., Schon, K., Tricarico, M. D., Swisher, J. D., Celone, K. A., and Stern, C. E. (2008). Working memory for social cues recruits orbitofrontal cortex and amygdala: a functional magnetic resonance imaging study of delayed matching to sample for emotional expressions. *J. Neurosci.* 28, 3718–3728.
Malle, B. F. (2004). *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction*. Cambridge, MA: MIT Press.
Miller, E. K. (2000). The PFC and cognitive control. *Nat. Rev. Neurosci.* 1, 59–65.
Miller, E. K., and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* 24, 167–202.
Miller, E. K., and Phelps, E. (2010). Preface: current opinion in neurobiology – cognitive neuroscience 2010. *Curr. Opin. Neurobiol.* 20, 1–2.
Moll, J., Krueger, F., Zahn, R., Pardini, M., de Oliveira-Souza, R., and Grafman, J. (2006). "Human fronto-mesolimbic networks guide decisions about charitable donation. *Proc. Natl. Acad. Sci. U.S.A.* 103, 15623–15628.
Monti, M. M., Osherson, D. N., Martinez, M. J., and Parsons, L. M. (2007). Functional neuroanatomy of deductive inference: a language-independent distributed network. *Neuroimage* 37, 1005–1016.
Murphy, G. L. (2002). *The Big Book of Concepts*. Cambridge, MA: MIT Press.
Murphy, G. L., and Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychol. Rev.* 92, 289–316.
Noveck, I. A., Goel, V., and Smith, K. W. (2004). The neural basis of conditional reasoning with arbitrary content. *Cortex* 40, 613–622.
Owen, A. M., McMillan, K. M., Laird, A. R., and Bullmore, E. (2005). N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies. *Hum. Brain Mapp.* 25, 46–59.
Pandya, D. N., and Barnes, C. L. (1987). "Architecture and connections of the frontal lobe," in *The Frontal Lobes Revisited*, ed. E. Perecman (New York: The IRBN Press), 41–72.
Patterson, R., and Barbey, A. K. (in press). "Causal simulation theory: an integrative cognitive neuroscience framework for causal reasoning," in *The Neural Representation of Belief Systems*, eds J. Grafman and F. Krueger (New York, NY: Psychological Press).
Pessoa, L. (2008). On the relationship between emotion and cognition. *Nat. Rev. Neurosci.* 9, 148–158.
Petrides, M. (2005). Lateral prefrontal cortex: architectonic and functional organization. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360, 781–795.
Prehn, K., Wartenburger, I., Meriau, K., Scheibe, C., Goodenough, O. R., Villringer, A., van der Meer, E., and Heekeren, H. R. (2008). Individual differences in moral judgment competence influence neural correlates of socio-normative judgments. *Soc. Cogn. Affect. Neurosci.* 3, 33–46.
Ramnani, N., and Owen, A. M. (2004). Anterior PFC: insights into function from anatomy and neuroimaging. *Nat. Rev. Neurosci.* 5, 184–194.
Rilling, J. K., Goldsmith, D. R., Glenn, A. L., Jairam, M. R., Elfenbein, H. A., Dagenais, J. E., Murdock, C. D., and Pagnoni, G. (2008). The neural correlates of the affective response to unreciprocated cooperation. *Neuropsychologia* 465, 1256–1266.
Rochat, P. (2009). *Others in Mind – Social Origins of Self-Consciousness*. New York, NY: Cambridge University Press.
Rolls, E. T., Hornak, J., Wade, D., and McGrath, J. (1994). Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. *J. Neurol. Neurosurg. Psychiatr.* 57, 1518–1524.
Ruby, P., and Decety, J. (2004). How would you feel versus how do you think she would feel? A neuroimaging study of perspective-taking with social emotions. *J. Neurosci.* 16, 988–999.
Salmon, W. (1998). *Causality and Explanation*. New York: Oxford University Press.
Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., and Cohen, J. D. (2003). The neural basis of decision making in the ultimatum game. *Science* 300, 1755–1758.
Simmons, W. K., Reddish, M., Bellgowan, P. S. F., and Martin, A. (2010). The selectivity and functional connectivity of the anterior temporal lobes. *Cereb. Cortex* 20, 813–825.
Sloman, S. A., Barbey, A. K., and Hotaling, J. (2009). A causal model theory of the meaning of "cause," "enable," and "prevent." *Cogn. Sci.* 33, 21–50.
Smith, R., Keramatian, K., and Christoff, K. (2007). Localizing the rostrolateral PFC at the individual level. *Neuroimage* 36, 1387–1396.
Stone, V., Cosmides, L., Tooby, J., Kroll, N., and Knight, R. (2002). Selective impairment of reasoning about social exchange in a patient with bilateral limbic system damage. *Proc. Natl. Acad. Sci. U.S.A.* 99, 11531–11536.
Tenenbaum, J. B., Griffiths, T., and Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends Cogn. Sci. (Regul. Ed.)* 10, 309–318.
Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *Monist* 59, 204–217.
Uttich, K., and Lombrozo, T. (2010). Norms inform mental state ascriptions: a rational explanation for the side-effect effect. *Cognition* 116, 87–100.
Wager, T. D., Jonides, J., and Reading, S. (2004). Neuroimaging studies of shifting attention: a meta-analysis. *Neuroimage* 22, 1679–1693.
Wager, T. D., and Smith, E. E. (2003). Neuroimaging studies of working memory: a meta-analysis. *Cogn. Affect. Behav. Neurosci.* 3, 255–274.
Weissman, D. H., Perkins, A. S., and Woldorff, M. G. (2008). Cognitive control in social situations: a role for the dorsolateral PFC. *Neuroimage* 40, 955–962
Wellman, H. M., and Liu, D. (2007). "Causal reasoning as informed by the early development of explanations," in *Causal Learning: Psychology, Philosophy, and Computation*, eds A. Gopnik and L. E. Shultz (New York, NY: Oxford University Press), 261–279.
Wolff, P., Barbey, A. K., and Hausknecht, M. (2010). For want of a nail: how absences cause events. *J. Exp. Psychol. Gen.* 2, 191–221.
Wood, J., and Grafman, J. (2003). Human prefrontal cortex: processing and representational perspective. *Nat. Rev. Neurosci.* 4, 139–147.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Received: 11 August 2010; accepted: 28 June 2011; published online: 27 July 2011.*
*Citation: Barbey AK and Patterson R (2011) Architecture of explanatory inference in the human prefrontal cortex. Front. Psychology* ***2****:162. doi: 10.3389/fpsyg.2011.00162*
*This article was submitted to Frontiers in Cognition, a specialty of Frontiers in Psychology.*