3. A Cognitive Neuroscience Framework for Causal Reasoning

Richard Patterson[1] & Aron K. Barbey[2]

Emory University, USA[1] and University of Illinois at Urbana-Champaign, USA[2]

Corresponding Author: Aron K. Barbey, Decision Neuroscience Laboratory, College of Applied Health Sciences, University of Illinois at Urbana-Champaign, 110 Huff Hall 1206 South Fourth Street Champaign, IL 61820, Barbey@Illinois.edu

Introduction

The ability to acquire, maintain, and utilize an up-to-date system of causal information and to infer effects from potential causes is essential for perception of our environment and for successful interaction with it. Thus causal perception and inference are tightly intertwined with virtually every sort of human cognition. However, there do exist at least three well-developed psychological theories that attempt to define and investigate causal inference as such—one representing causes as forces and inferring causal conclusions by use of *Force Composition* (FC; Barbey & Wolff, 2006, 2007, submitted; Wolff, Barbey & Hausknecht, 2010), another representing causal relations and inferences via Baysian Net diagrams, *Causal Model Theory*, (CM; Sloman, Barbey & Hotaling, 2008), and a third based on more abstract models representing types of possible situations in which given causal premises and conclusions are true or false, the *Mental Models* framework (MM, Goldvarg & Johnson-Laird, 2001; Johnson-Laird & Goldvarg-Steingold, 2007).

Section II presents the essentials of these three theories and their main neural implications. But we believe that none of the three provides a comprehensive account of causal inference, or even captures the single most central—and evolutionarily oldest—mode of causal representation and reasoning, namely, the constructing and "running" of causal scenarios, based on current perception of specific objects, agents, and events, awareness of the situations in which these factors are embedded, and relevant background information derived from past experience. *Causal Simulation Theory* (CS) falls within the domain of theories of "embodied cognition" in which beliefs and inferences are represented and instantiated neurally in essentially the same broadly perceptual manner as sense perception, proprioception, et al. (Barsalou, 1997).

Nonetheless a comprehensive account of causal reasoning must be pluralistic. In the first place, CS alone is far from the whole story about causal inference, and in some situations may play only a minor role, or no role at all. Secondly, human (causal) reasoning is flexible, resourceful, opportunistic, even if sometimes shallow, muddled and confused: it calls upon awhatever means might promise to be helpful, and often uses different processes in sequence or in tandem. CS in particular joins forces naturally and effortlessly in a great many contexts with the processes postulated in FC, CM, and MM. In addition, language will in many situations contribute to the shaping and manipulation of simulations. Finally, rule use will be critical in some situations, whether in the form of rules of thumb, ad hoc heuristics, or causal-inferential schemata assumed to be fundamental and law-like. So although causal simulation is the single most common style of human causal inference (and may be the default mode), causal reasoning often draws on and combines multiple psychological abilities and neural systems, and can be accounted for only by a pluralistic account that describes the various processes that are dynamically recruited in particular situations, characterizes the types of reasoning context in which human beings are likely to use some inferential process or combination of processes—and identifies the neural systems that implement such reasoning.

Section III reviews the relevant neuroscientific literature on causal perception and inference. Although there are some extremely interesting results to report, this work is not designed to test any of the four specific approaches to causal inference under consideration here, and justifies only limited and preliminary judgments about the merits of those theories. Speaking more generally, research in this area is at an early, exploratory stage, so we will close with a number of suggestions for future work.

Four Psychological Theories of Causal Inference and Their Neural Implications

Force composition

The *Force Composition Model* (FC; Barbey & Wolff, 2006, 2007, submitted; Wolff et al., 2010) extends an earlier theory of force dynamics (Wolff, 2007), itself derived from Leonard Talmy's theory of force dynamics (Talmy, 1988). The theory pertains to the representation of causal relations (e.g.,, *A causes B*, *B prevents C*) and how people combine representations to draw new causal conclusions. FC holds that people conceive of causal relations – physical, psychological, social, etc. – in terms of force, and reason about them using representations that reflect configurations of forces. Although *force* is not defined in the theory, it appears to include all causal factors that have both magnitude and direction, and that can combine with other such factors to result in some outcome. We are skeptical about the ability of FC to represent all causal reasoning, because causal reasoning is not always concerned with forces, and even where causal forces are involved it is not always clear that these can be suitably composed to give a result. But we defer further discussion of these issues, since one need not claim that FC is *the* theory of causal inference in order to recognize its potential importance as part of a more pluralistic approach.

FC analyses the main types of causal relation in terms of a Patient (*P*), an Affector (*A*), and an Endstate (*E*), plus Tendencies of the Patient and Affector toward or away from the Endstate. Distinct combinations of tendencies and outcomes define the relations of *Cause*, *Help* / *Allow*, and *Prevent.* Figure 1 shows the configuration of forces underlying these causal relations.

[Insert Figure 1 about here]

Figure 2 shows how configurations of force can combine to generate conclusions, where the affector in the conclusion is the affector from the first premise (**A**); the end state in the conclusion is the endstate from the previous premise (**E**); and the patient in the conclusion is the resultant of the patient vectors in the premises (**B** + **C**) (for further detail, see Wolff et al., 2010).

[Insert Figure 2 about here]

The authors do not say exactly how people represent forces. Some of the most interesting supporting experiments use simplified visual simulations of causal scenarios (e.g.,, images of bumper cars colliding, fans blowing toy sailboats, pedestrians crossing a street as directed by a policeman). The results of these experiments are therefore consistent with the possibility that participants use causal simulations (at least in part) to infer outcomes. However, Barbey and Wolf (2006, 2007, submitted), and Wolff et al. (2010) characterize the reasoning involved in the more abstract terms of "compositional force diagrams". We suggest that such diagrams might serve either (a) as the sole vehicle of causal inference (as the articles seem to suggest), or (b) in a side-by-side manner with perceptually-based simulations wherein reasoners "toggle" back and forth between the two, or (c) by superimposition of causal force arrows on causal simulations. We know of no psychological experimentation testing for these different possibilities.

Neural Predictions of Force Composition Theory

If FC diagrams function in a stand-alone manner, one should expect activation in neural systems central to visual-spatial tasks such as maintenance, monitoring, and manipulation of spatial representations. This suggests a common causal reasoning process, at least for the bulk of causal reasoning, one recruiting occipital (BA 17, 18, 19) and parietal (BA 7, 39, 40) regions involved in the construction and manipulation of graphic materials (e.g.,, Palmer, 1999; Koenigs et al., 2009), in addition to regions involved in the representation of force, such as motor (BA 4) and somatosensory cortices (BA 3, 1, 2; Figure 3). Where FC diagrams function in conjunction with simulations, there should be activation in visual-spatial systems plus other systems, depending on the kinds of forces and causal relata involved (physical and visible, psychological, economic, etc.). Thus the FC model motivates a modified form of the *domain-general reasoning hypothesis*, according to which causal reasoning will in general recruit areas involved in the formation and manipulation of force vectors, but supplementing this with different patterns of modality-specific activation based on the type of forces involved.

[Insert Figure 3 about here]

Causal Model Theory

The *Causal Model Theory* (CM) of the meaning of *CAUSE*, *ENABLE*, and *PREVENT* (Sloman, Barbey, & Hotaling, 2009; Chaigneau & Barbey, 2008) utilizes the graphical formalism of causal Bayes nets to represent and make inferences about causal relations (Pearl, 2000; Spirtes, et al., 1993; for a non-technical introduction see Sloman, 2005). The critical idea is that of a *link*: a link between *X* and *Y* represents a causal mechanism that has *X* as one of its inputs and *Y* as the output, and is defined in terms of *intervention* (Woodward, 2003): a causal path involving one or more links exists between *X* and *Y* if intervening on *X* changes or would change the value of *Y* (and not the converse). CM makes little use of the technical apparatus of the causal Bayes nets approach, operating instead with *qualitative* (diagrammatic) representations of causal models. In general, two events *A* and *B* could have any number of assumed causes and effects, all of which one represents as directional arrows or edges in a causal net. In the end CM is notably abstract, for although it does speak of "mechanisms*"*, these turn out to be simply functions from inputs to outputs, with the proviso that the latter cannot temporally precede the former.

A system of causal links in a Bayes net corresponds in turn to a set of structural equations, in accordance with the rule that effects are a joint function of all their causes. For instance, the causal model:

[Insert Figure 4 about here]

expresses the structural equation: B:=f(A, ε), where ε represents uncertainty due to other variables not represented in the model. The possibility of uncertainty allows the relation between A and B to be probabilistic, even if all actual causes involved are deterministic (Figure 4). Figure 5 summarizes the structural equations underlying *CAUSE*, *ENABLE*, and *PREVENT* according to CM theory.

[Insert Figure 5 about here]

Equations expressing causal relations can then serve as premises in order to derive, in conjunction with certain "Processing Assumptions", causal conclusions. For example, *A causes B* and *B causes C* combine to give the conclusion *A causes C*, as follows:

B := A (1)

C := B (2)

C :=A (By substitution of (1) into (2), as stated by a Processing Assumption).

As with FC, there are different ways in which people might represent causal information in the form of premises or conclusions, and the theory does not clearly say which describes actual reasoning. One might use just spatial Bayes net diagrams, or causal equations, or both together, depending on the situation. We know of no empirical work testing for these alternatives, so although the use of diagrams would seem the most plausible option, this is at the moment still an open question. Also, we find some significant problems with the manner in which these equations, supplemented by certain Processing Assumptions are supposed to generate conclusions from causal premises. But again, we defer that discussion for the time being, because the basic framework is viable and, we think, needed for a comprehensive account of causal reasoning.

Neural Predictions of the Causal Models Theory

First, manipulation of causal graphs or diagrams will, recruit a broadly distributed network of neural systems, including visual-spatial processing regions in the occipital (BA 17, 18, 19) and parietal (BA 7, 39, 40) lobes. By contrast, logical operations (those involved in deduction, and in manipulation of structural equations in accordance with processing rules) are subserved primarily by regions of the prefrontal cortex and particularly lateral (BA 8, 9, 46, 44, 47) and orbitofrontal prefrontal cortex subregions (BA 10, 11) (Figure 6; for a review, see Goel, 2007). Further, both sorts of areas will be involved if graphs and equations are both involved in some episodes of causal reasoning. CM theory thus supports the *domain-general reasoning hypothesis*, but potentially in two different ways: causal reasoning either constructs or manipulates domain-general graphical representations (graphical causal models), or domain-general structural equations and processing rules, or both. In all cases it recruits systems from this set of options across different causal reasoning tasks and materials. Note finally that the theory should have something to say about the conditions under which one should expect one or another of these processes to come into play.

[Insert Figure 6 about here]

Mental Model Theory

Mental Model Theory (MM) “purports to solve three puzzles: first, what causal relations mean; second, how they are mentally represented; and, third, how people make inferences from them” (Goldvarg & Johnson-Laird, 2001, p. 566; see also Johnson-Laird & Goldvarg-Steingold, 2007). “Each model corresponds to a possibility, and models are labeled to distinguish physical, deontic and logical possibilities…[b] the structure and content of a model capture what is common to the different ways in which the possibility can occur…[c] naive reasoners imagine the states of affairs described by premises; they construct mental models of them, and they establish the validity of an inference by checking whether its conclusion holds in these models (Goldvarg & Johnson-Laird, 2001, p. 566).” In order to minimize the load on working memory, people represent as little information as possible, with the result that mental models are often not complete. Table 1 summarizes the mental models underlying CAUSE, ALLOW, and PREVENT relations.

[Insert Table 1 about here]

In causal inference one does not use inferential schemata or rules of inference, but generates mental models based on the premises, then checks for validity by determining whether a given conclusion is true in all the mental models of the premises (valid), or whether there can be a counterexample to the inference (a possible situation in which all the premises are true but the conclusion false). Table 2 illustrates how a conclusion follows from the premises *A causes B*, and *B prevents C* according to the mental model framework.

[Insert Table 2 about here]

On this theory on can make probabilistic judgments as well, since with an invalid inference one can judge what proportion of the possible mental models of the conclusion are consistent with the premises (e.g.,, 3 of 4), and use this as an estimate of probability.

The mental models approach constitutes a further variety of domain-general inference procedure. The theory suggests an abstract spatial (but not necessarily visual, Johnson-Laird, 1998) representation of types of possible situation, with the possibility of combining mental models to infer causal conclusions. There is some quite interesting behavioral evidence for the theory, although we believe that there are also some theoretical problems that still need to be addressed, and that the experimental materials and task demands are too restricted in key respects to support any claim for mental models as a comprehensive account of causal inference. But the theory is substantial and economical, and probably does correctly capture the manner in which humans draw causal conclusions in certain situations.

Neural Predictions of Mental Model Theory

Given the proposed role of mental models in both deductive and inductive reasoning, this theory predicts that these forms of reasoning will recruit common neural systems, and that these systems will be engaged in all types of causal reasoning. Johnson-Laird (1995) further claims, "The model theory also makes a critical prediction about the role of the cerebral hemispheres in reasoning. As Whitaker et al. (1991) first noted, the construction of models is likely to depend on the right hemisphere." Thus MM theory predicts that deductive and inductive causal reasoning will both (1) primarily recruit right hemispheric regions. Furthermore, to the extent that a search for counterexamples is performed, the (2) right frontal pole (BA 11), which has been implicated in evaluative reasoning (Kroger et al., 2008), should also be recruited (Johnson-Laird, personal communication). The proposed spatial nature of mental models motivates the prediction that (3) occipital (BA 17, 18, 19) and parietal (BA 7, 39, 40) regions implicated in visual-spatial processing will be engaged (Figure 7; Knauff et al., 2002). Finally, MM supports the *domain-general reasoning hypothesis*, according to which causal reasoning recruits domain-general cognitive representations (mental models) and will therefore engage common neural systems across different causal reasoning tasks and materials.

Causal Simulation Theory

Causal Simulation Theory (CS) remains close to the perceptual experiences in which we apprehend causal events, maintaining that most causal reasoning takes place via simulation, in broadly perceptual terms, of causal scenarios. A large body of neuroscientific and psychological evidence indicates that humans represent themselves and their world by building up multimodal, hierarchically organized perceptual and experiential wholes (such as percepts of complete objects, situations incorporating objects, agents, and events, egocentric and allocentric maps of spatial environments, etc.) starting from smaller parts ranging down to basic modality-specific items responded to by “feature detectors” (edges, tones, movements, orientations, and so on). In a broad sense of ‘perceptual’, including feelings, sensations, motions, emotions, etc., the resulting representations are “perceptual symbols” (Barsalou, 1997), organized in step-wise fashion through levels of increasing complexity (see Damasio, 1989 on the role of “convergence zones” in this process).

Perceptual symbols are not in general linguaform. And although our view is logically neutral on the question of the ultimate nature of causal beliefs, it clearly harmonizes well with the Churchlands' "first pass" account of how multi-dimentional neural states (as opposed to linguaform representations), and in particular, neural maps, can track, represent, or be about, items in one's environment so as to guide everyday navigation through the world. Their approach builds both on work concerning the modeling of neural nets and psychological processes in connectionist networks, and on the now well-established presence of numerous homomorphic maps in the brain. We agree that beliefs and other traditional mental states are not necessarily "linguaform", as thought by many (but certainly not all) philosophers. So aside from the rear guard action against certain philosophers, the big problem—or better, the exciting prospect—is the further working out of a theory according to which neural states can themselves represent or be about items in the "external" environment and constitute an effective map of the causal properties and histories of the things among which we live. At the same time we are somewhat more sympathetic to linguaform representations than the Churchlands appear to be, since we regard them not only as “an extremely subtle and useful means of communicating representations among humans” (Churchlands, 2010) but, once language is up and running, as extremely useful instruments for the creation, manipulation, and application of one's own non-linguaform representations—and possibly even as essential parts, along with non-linguaform constituents, of some beliefs. So we anticipate in any event further study of the combination and co-operation of linguaform representations with other types of representation, and not *necessarily* the scientific withering away of linguaform beliefs.

Our representations of things and agents, then, with their causal powers and liabilities, along with the larger environments in which these items are situated and their causal histories, are all largely perceptual in nature. That is, the representations enlisted in memory, fantasy, planning and execution of actions, and so on, are predominately of the same basic sort as those generated in everyday perceptual experience. At its outer reaches—i.e., abstract concepts such as *exclusive disjunction* or *freedom*, the theory of Perceptual Symbol Systems is controversial and still under development. Our Causal Simulation theory is neutral on such issues, since it does not insist that all causal thought relies on simulations. However, we note below some important ways in which even very abstract concepts and thoughts are routinely represented via perceptual representations and simulations.

In particular situations, then, we can call up from memory, or generate “on the fly”, simulations of an actual or potential causal scenario and “run” the simulation so as to generate an array of possible developments of that situation and anticipate possible responses to it. Or we can envision possible antecedent scenarios that might have caused, and might explain, a given situation. This basic picture needs to be elaborated in detail and critically evaluated both theoretically and experimentally—and the latter both behaviorally and neurologically. As for the elaboration, we set forth here the main points of the theory.

1. Simulations can be "off the shelf" or "stock" items, like commonly experienced schemata, scripts, and narratives, or created "on the fly" (like ad hoc categories, Barsalou, 1983) and recruited dynamically to meet the needs of the moment.
2. Since every moment has its situated needs and purposes—for actual response, anticipation and preparedness, understanding, etc.—people probably run small-scale simulations continuously, where these are nested within scenarios of greater generality and longer time span, and where simulations are constantly updated or corrected, especially at more "local" levels (Zacks, personal communication).
3. Simulations can be partial, schematic, and simplified, or relatively complete and detailed.
4. One can run simulations in a continuous and holistic way, or piecemeal and in a "stop and go" fashion—as when imagining the end effect of the movements of a series of interlocking gears, or pulleys and wheels (Hegardy, 2004).
5. Simulations can involve any combination of external or internal perceptual modalities – visual, auditory, emotional, social, haptic, spatial, kinetic, visceral, etc. (for a review of simulation mechanisms in social information processing, see Barbey & Grafman, in press a, b; Barbey et al., 2009 a).
6. Simulations can facilitate counterfactual causal reasoning as well as reasoning about what might happen next given some actual situation (for recent reviews, see Barbey et al., in press, 2009 b).
7. The intricate processes involved in producing the full range of simulations are not well understood, but they are not confined to "brick and mortar" procedures (the lowest level "bricks" being elementary perceptual features): for example, humans readily devise metaphors and analogies (Gentner & Colhoun, in press), and conceptual "blends" of various sorts (Fauconnier & Turner, 2003), and put these to use in causal reasoning.

8. Perceptual scenarios can represent causally significant abstract things and properties, or “theoretical entities”, such as viscosity, compassion, or subatomic forces (Schwartz, 1999).
9. Perceptually based simulations, even those of particular objects or situations, can represent general *types* of things, situations, or events, depending on how one uses the simulation in one’s reasoning – above all, on whether or not one uses only features of the particular item that are shared by all members of a given type or category (Barsalou, 1997, Aristotle in Barnes, 1991).
10. Language will play an important role in human causal reasoning, frequently guiding the construction, interpretation, and manipulation of a simulation, and the focusing of attention over time on relevant features of a simulation. (For evidence that “text-based” causal reasoning involves simulations, see Kurby et al., 2008). Especially important is the power of language to build into a simulation many factors that are not directly sensory, but that are causally relevant – e.g.,, density or viscosity in cases of “mechanical” causation, intentions and plans of a sentient agent, the “industrial power” of a nation, and so on. A relatively simple representation can integrate and signify a great deal of verbally presented information about causal properties whose effects are then manifested in the running of simulations.
11. Causal simulations can be accurate and insightful, or superficial and inaccurate. People often believe they know how a thing works, and even think they can “picture” this consciously, when in fact their understanding is quite shallow, loose, or erroneous. (Rozenblit & Keil, 2002).
12. Causal simulations can function in a “stand alone” manner, but simulations readily combine with or support the use of language, rules, composition of forces, Bayes nets (including Power PC calculations), and mental models.

In sum, people draw on beliefs about the causal powers and histories of things, combining this with current perceptual information in a particular life situation to call up from memory, or to construct on the spot, appropriate causal scenario(s), then run these simulated scenarios—if appropriate, in conjunction with other "inferential aids" such as language, rules of thumb, force diagrams, etc.--in order to generate possible further developments and possible responses to these, or to gain causal understanding of a situation. Current causal simulations are custom tailored from the materials at hand, to suit one's particular needs and purposes.

Neural Predictions of Causal Simulation Theory

CS theory provides a framework for representing causal knowledge in the form of perceptual simulations, and predicts that causal reasoning will recruit (1) a broadly distributed system of modality-specific brain areas and (2) regions that are involved in constructing perceptually-based simulations of past and future events, which include the medial prefrontal cortex, the precuneus and retrosplenial cortex, and regions of the medial and lateral temporal cortex (for a review, see Schacter et al., 2007). Figure 8 illustrates the brain mechanisms underlying the simulation of event knowledge according to CS theory.

[Insert Figure 8 about here]

CS theory motivates the *domain-specific reasoning hypothesis* (Barbey & Barsalou, 2009). According to this hypothesis, the neural areas underlying a particular type of reasoning, such as causal inference, may vary widely and show little in common as the specific materials and tasks vary. Because different materials and tasks produce different patterns of modality-specific activation, there may be no areas of neural activity common to all sorts of causal reasoning. However, some areas may be recruited much more frequently than others, and some may be common to many instances of causal reasoning. This is to be expected not because causal reasoning uses a domain-general procedure, but because some systems, such as those supporting creation and manipulation of spatial or visual representations are important for numerous different kinds of causal situations (psychological, physical, social, et al.). Similarly, neural systems for language processing need not always, but very often will be, involved, e.g.,, in many learning situations, in guiding the construction of simulations, and in focusing attention on relevant aspects of a simulation. The main point is that the patterns of activation will vary in a manner that reflects both the centrality of different perceptual modalities in thinking about different sorts of materials, and the recruitment of language, rules, or any of the four styles of inference described here.

[Insert Table 3 about here]

Review of Neuroscience Literature

Causal Reasoning

To our knowledge there are relatively few brain imaging studies that directly address causal reasoning. Some of these attempt to isolate features of causal cognition as such via carefully constructed behavioral experiments and then identify (usually using functional Magnetic Resonance Imaging, fMRI) the underlying neural systems, whereas others in effect study causal reasoning as a special case of some more general type(s) of cognition—e.g.,, reasoning with familiar versus unfamiliar material, reasoning with materials that do or do not conflict with one's previous beliefs, reasoning deductively or inductively. At this point there is some, but not much, basis for evaluating the relative merits of the four theories under consideration here.

Fonlupt (2003) evaluated the neural systems supporting perception of mechanical causation in Michotte's classic launching events (Michotte, 1963), wherein a ball travels horizontally across a computer screen and collides with a ball located in the center. The apparent collision results in the second ball *launching* away from the first horizontally across the screen and elicits the perception that the first ball caused the second to move (Michotte, 1963). Fonlupt compared the neural response produced by the launching event to that elicited by control events in which the first ball passes below the second ball without a collision (*non-causal condition*). Of primary interest were the neural systems engaged when subjects judged either (i) the presence or absence of causation, versus (ii) the direction of the ball's motion. Fonlupt observed a reliable increase in medial prefrontal cortex (BA 11) activation in judgments of causality relative to judgments of ball movement. Moreover, this increase occurred during both the causal and non-causal conditions, suggesting that the signal increase was specifically associated with the process of making a causal judgment and not with the perception of actual causality, or the making of a perceptual judgment (e.g.,, of direction of motion).

Roger et al., (2005) investigated whether causal perception and causal inference rely on common or distinct hemispheric regions. The authors tested two callosotomy (split-brain) patients and a group of neurologically intact patients. Of primary interest was assessing the role of the left versus right hemispheres in (i) the perception of causal events (i.e., Michotte's launching event) and (ii) causal inference tasks where the relation between a candidate cause and observed effect must be inferred from simple covariations rather than perceived.

Roger et al. found that the perception of causality and causal inference from covariations depend on different hemispheres of the divided brain: whereas the perception of causality relied on the right hemisphere, causal inference engaged the left hemisphere. They add, however, that in the intact brain both hemispheres may be involved in both sorts of causal cognition (as in fact occurred in normal subjects), but that one hemisphere might "jumpstart" the relevant process.

Since the causal inference involved in this experimental setup depends on detection of simple covariational information, it is consistent with the Causal Model (Bayes Net) approach, as the authors note (Roser et al., 2005). It does not show, however, that in other experimental or "real" situations causal reasoning would not use force composition, mental models, or causal simulations, or that these other sorts of reasoning process would show similar hemispheric dependence.

Fugelsang and Dunbar (2004) investigated brain states underlying comparatively complex causal reasoning by presenting subjects with causal theories, then covariational data, and varying both the plausibility of the theory (judged by the plausibility of the theory's proposed causal mechanism) and the consistency/inconsistency of the data with the given theory. When individuals reasoned with evidence that was consistent with existing causal beliefs, a network of brain regions widely associated with learning and memory was engaged, including the caudate and the parahippocampal gyrus. By contrast, evaluating data *inconsistent* with a *plausible* theory resulted in a pattern of activation widely associated with error detection and conflict resolution, including the anterior cingulate cortex (BA 24/32), posterior cingulate, and the precuneus (BA 7). Put another way, "the basic finding is that people weight the covariation-based evidence stronger when it follows from a theory that contains a plausible mechanism of action than when the evidence follows from a theory that contains an implausible mechanism of action" (see Fugelsang & Thompson 2000, 2003; Fugelsang & Dunbar 2004). Fugelsang and Dunbar propose that people's beliefs and expectations "act as a biological filter during evidence evaluation by selectively recruiting learning mechanisms for evidence that is consistent with their beliefs and error detection mechanisms for evidence that is inconsistent with their beliefs" (p. 1752). These researchers do not claim to have isolated the neural underpinnings of causal reasoning as such; rather, they argue that reasoning about causal questions—as when weighing evidence in a court of law—does in fact follow a pattern previously observed in evaluation of information that is or is not consistent with previous beliefs, and that this has potentially important practical implications for how one might in a courtroom context try to overcome unconscious bias in the evaluation of evidence.

Deductive Reasoning

Brain imaging studies of deduction are especially important for MM theory, but have application also to CS (which, again, countenances the co-operation of simulation with MMs). Here we offer only a very brief overview of key studies and their bearing on current theories of causal reasoning.

The MM theory clearly gives a central role to deductive reasoning from causal premises to causal conclusions, and provides for "free", as it were, an account of inductive causal inference. Note, however, that this is not inference by logical rule, but inference by search for counterexamples. (The latter is sometimes called a "semantic", the former a "syntactic" method of proof.) The role (if any) for deductive reasoning in the FC and CM theories is less clear. CS allows that under appropriate conditions deductive or inductive processes may be involved, and that they may sometimes (as in MM) be the sole process involved. But it also predicts that very frequently humans will use CS as at least a support for deduction: that is, one will carry out either semantically – or syntactically-based inference with the help of simulations of the kinds of relata involved. This amounts to taking a particular item as an "arbitrary" instance of a category, so that one's conclusions about that instance apply to all members of the category. (See Section II.D above, Point 9).

Table 4 reports the results from a recent review of the neuroscience literature on deductive reasoning conducted by Goel (2007). The observed recruitment of broadly distributed neural systems – spanning the frontal, temporal, and occipital lobes – indicates that the neural architecture of deductive reasoning is highly sensitive to the content of the reasoning problem (concrete/abstract, logical/belief-biased, familiar/unfamiliar, etc.) and processing demands of the reasoning task (transitivity, categorical syllogisms, conditionals, etc.).

In general, neural studies on processes of deduction and induction will help evaluate or constrain theories of causal inference only to the extent that such theories include a role for those processes. MM gives deduction a central role, but in most theories its role is rather marginal. Still, deduction is an important resource and will contribute to actual causal reasoning when the situation calls for it. We make some suggestions below about how theories of causal inference might make greater use of modern formal systems of logic.

[Insert Table 4 about here]

Conclusions

First, the FC, CM, and MM frameworks are well suited for particular domains and circumstances of causal inference and each captures important aspects of causal knowledge. Our overview, however, identifies significant limitations of each framework and suggests that these theories are insufficient to account for the diversity of causal beliefs and the variety of inferences these representations support. Second, we develop a broader, more pluralistic theory of causal reasoning centering around causal simulations (see Secs. II.D) but incorporating their ready potential for supporting and interacting with other processes. Third, we motivate neurobiological predictions of each theory and review the neuroscience literature on causal and deductive reasoning, providing evidence to confirm the predictions of the CS framework. Our findings suggest that the neural systems underlying reasoning are highly sensitive to the content of the reasoning problem (consistent/inconsistent with prior beliefs, familiar/unfamiliar, etc.) and the form of inference required to draw a conclusion (deduction, causal inference). The CS theory receives support because it predicts the presence of a broadly distributed neural architecture for human reasoning and the selective recruitment of modality-specific cortices based on the content of the reasoning problem and the type of inference involved. This framework also receives support from the many studies that exhibit effects of task specificity and cognitive demand (Table 4). As task conditions change, so do the neural systems that represent and process the relevant information. In general, then, causal reasoning tends to recruit broadly distributed and diverse neural systems, depending (at least) on type of task, degree of difficulty, and the nature of the materials involved.

Future Research

We close with some suggestions, theoretical and experimental, for future research focusing on the neural systems underlying causal cognition. First, all four theories of causal inference surveyed above enjoy substantial (behavioral) experimental support and are essential, from our pluralist point of view, to a comprehensive account of causal reasoning. However, all four have yet to be investigated from a neuroscientific point of view. To begin with, this will involve imaging studies using the kinds of materials and experimental instructions already devised for use in behavioral studies. Thus a first set of imaging experiments lies almost ready at hand. But we strongly suggest that results for all four theories be examined for differences in neural activity, and especially for connections among regions of activity as materials and instructions vary. From our point of view the overarching question is not, ‘Which theory (of the four) is correct?’, but ‘Which inferential process or combination of processes, and which underlying neural systems, operating in parallel or in sequence, are recruited to carry out causal inference, and under what sorts of circumstances?’

Second, the accounts of FC, CM, MM, and CS developed so far need to be pushed further (e.g.,, as noted earlier, with regard to the specific ways in which people represent causal information within a given framework), so that imaging experiments can test for (1) the neural activity predicted by different theories, and (2) the kind of multi-system activity predicted by domain-specific (or multimodal) approaches such as “pure” CS and the more comprehensive pluralistic account we propose here. Regarding CS in particular, imaging studies of causal *perception* and its possible differences from causal *inference* should build on the interesting work in Roger, et al. (2005) by looking at a variety of types of physical causality – crushing, bending, pulling, shattering, and other "naturalistic" or "ecologically relevant" phenomena in addition to launching events. Among other things, this should yield a better basis than we have now for evaluating theories about a unitary neural system for causal perception (whether or not it constitutes a classical module, and whether or not it is innate).

Third, CS in particular suggests the extension of research to other perceptual modalities than vision, and especially to audition. Many causal events make characteristic sounds—the cracking of a stick, crushing of a peanut, tearing of a piece of paper, smashing of glass, etc. Imaging studies could reveal whether there is overlap in neural processing of such events with those involved in visual perception of causality—which would be an obvious neural prediction of a unitary system for perception of physical causality.

Fourth, future research should build also on the work of Barbey and Wolf using schematic simulations of "real life" social situations to extend neural findings to reasoning about psychological and agent causality (for recent reviews, see Barbey & Grafman, in press a, b; Barbey et al., 2009 a). However, researchers should not shy away from using a combination of appropriate visual material with causally relevant verbal material (such as that used in text-based studies of causal processing). The latter could help control participants' use of background knowledge as well as their interpretation of visual materials. This in turn should make possible the use of experimental materials relevant, even in fairly subtle ways, to exploring the neural underpinnings of everyday causal interpretation in terms of agents' intentions, emotions, desires, and beliefs. This work would need to capitalize on recent important advances in our knowledge of the neural systems involved in emotion, decision making, and other psychological factors involved in causal perception, explanation, and inference.

Fifth, recent work on mirror neurons also needs to be brought into the discussion of causal cognition. The discovery of these neurons about twenty years ago has given a boost to philosophical proponents of "simulation theory" as opposed to "theory theory" concerning the manner in which people apprehend or understand the emotions, intentions, etc., of others (e.g.,, Iacoboni et al., 1999). So it may turn out that mirror neurons make an essential contribution to a great deal of human causal thought. Here future work should introduce the kinds of experimental protocols that could reveal patterns of mirror neuron activation—and co-activation or sequential activation with other neural systems—in response to different sorts of causal materials and tasks. It would be especially important to find out how identifiable populations of mirror neurons interact or co-operate with other neurons, and thereby to determine what precise role mirror neurons play in the many areas for which they seem at least potentially important.

Sixth, temporal sequencing of neural activity underlying causal cognition needs to be investigated as closely as possible. Much previous imaging research has of necessity used methods yielding a relatively high spatial resolution but low temporal resolution or vice-versa. Comparison of relatively high-resolution spatial and temporal data for performance of the same tasks by the same participants under the same circumstance—and ideally though simultaneous use of different imaging methods—and would be important, for example, in exploring the intriguing possibility of one hemisphere "jump starting" a process of causal reasoning which then involves both hemispheres (Roger et al., 2005). But the recommendation would apply to the evaluation of any theory—such as CS—that predicts the dynamic recruitment of different neural systems in causal reasoning.

Seventh, studies of the neural correlates of deduction and induction should take detailed account of the variety of potentially relevant logical systems available today. These are not in general intended by logicians to describe actual cognitive activity but they can, like formal theories of probability, decision making, or game theory, be useful sources of hypotheses about causal (and other) reasoning. Previous neural and psychological studies have used logical materials involving the basics of one or another logical system, usually categorical syllogistic or propositional logic in the case of deduction, but many other possibilities might be explored. For example, contemporary propositional logic uses “material implication” (where the truth value of ‘if *p*, then *q*’ depends only on the combination of truth values of *p* and *q*, and not on any conceptual or causal or other connection between *p* and *q*), whereas human reasoning usually assumes some sort of relevant connection between the antecedent and consequent of a conditional statement, and between premises and conclusion of an argument. In the case of causal reasoning this is especially obvious. Given that, why not look to existing work on relevance, causal, and modal logics (e.g.,, Barbey et al., 2009 a) for hypotheses about how humans might incorporate such concepts as *causality*, *causal necessity,* or *conceptual relevance* into deductive or other sorts of reasoning? Exploration along these lines might well contribute to construction of psychological and neural theories that reflect the flexibility and adaptability of human reasoning. Conversely, neural and behavioral evidence about deductive causal inference needs to be considered in light of a broader range of logical interpretations or models than has yet been explored.

Finally, and very programmatically, research should extend beyond the realm of specifically causal inference to that of other explanatory inferences. The study of diagrams in general would be of interest, partly because of their usefulness as aids to the useful organization of explanatory information of many types—in geometry (but also in number theory, and general proportion theory, as demonstrated already in Euclid), in engineering (circuit diagrams, “fishbone” diagrams), logic (Venn diagrams, Euler Circles, Pierce’s deductive schemata, Frege’s Begriffschrift or “conceptual notation”, semantic tableaux), tree structures in “intuitive biology”, not to mention Force Compositon and Bayes Net Diagrams. Diagrams may or may not form a “natural kind” within the larger domain of explanatory symbol systems, but certainly the widespread effectiveness of spatial diagrams not just as aids to memory but to understanding merits investigation. Beyond that one could look even more broadly at visual-spatial explanations in general – e.g.,, steps in a magic trick as pictured in an instruction book; illustrated stages of descent via alcoholism from “Happy Home” to “Pauper’s Grave”; drawings that explain how undertows form, and so on. Again, the larger project would be to investigate how the mind seeks explanatory understanding, and in particular how it selects and then manipulates different modes of representation, often dynamically recruiting an appropriate mix of these elements, to arrive at understanding under various sorts of conditions. Finally, do all (or many) varieties of understanding really have anything in common besides the name? If so, is there at some very high level of integration an identifiable even if distributed neural system for explanatory understanding? Is there even a general inborn human drive for explanation (Gopnik & Glymour, 2002), of which an impulse to construct causal explanations are one manifestation?

In short, look more closely at each and every tree, but think also about the nature and extent of the forest, and the yet more inclusive ecosystem to which these might belong.

References

Acuna, B.D., Eliassen, J.C., Donoghue, J.P. & Sanes, J.N. (2002). Frontal and parietal lobe activation during transitive inference in humans. *Cerebral Cortex, 12,* 1312-1321.

Barbey A.K. & Barsalou L.W. (2009) Reasoning and Problem Solving: Models. In L. Squire (Ed.) *Encyclopedia of Neuroscience*, Vol. 8 (pp. 35-43). Oxford: Academic Press.

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

Barbey, A.K. & Grafman, J. (in press b). An integrative cognitive neuroscience theory for social reasoning and moral judgment. *Wiley Interdisciplinary Reviews: Cognitive Science.*

Barbey, A.K., Krueger, F. & Grafman, J. (in press). Structured event complexes and mental models for counterfactual inference. In M. Bar (Ed.), *Predictions in the Brain: Using our Past to Prepare for the Future.* Oxford University Press.

Barbey, A.K., Krueger, F. & Grafman, J. (2009 a). An evolutionarily adaptive neural architecture for social reasoning. *Trends in Neuroscience*s, 32, 603-610.

Barbey, A.K., Krueger, F. & Grafman, J. (2009 b). Structured event complexes in the prefrontal cortex support counterfactual representations for future planning. *Philosophical Transactions of the Royal Society of London: Biological Sciences*, 364, 1291-1300.

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

Barbey, A.K. & Wolff, P. (submitted). Composing causal relations in force dynamics.

Barbey, A.K. & Wolff, P. (2007). Learning causal structure from reasoning. In D.S. McNamara & J.G. Trafton (Eds.) *Proceedings of the 29th Annual Conference of the Cognitive Science Society*. (p.713-718). Mahwah, NJ: Erlbaum Press.

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

Barnes, J. (1991). *The Complete Works of Aristotle.* Princeton Univ. Press.

Barsalou, L.W., Simmons, W.K., Barbey, A.K. & Wilson, C.D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences, 7,* 84-91.

Barwise, K. J. & Etchemendy, J. (2002). *Language, Proof and Logic*. Seven Bridges Press, New York: NY.

Bottini, G., Corcoran, R., Sterzi, R., Paulesu, E., Schenone, P., Scarpa, P., Frackowiakk, R. S. J. & Frith, C., D. (1994). The role of the right hemisphere in the interpretation of figurative aspects of language: A positron emission tomography activation study. *Brain, 117,* 1241-1253.

Brownell, H. H., Simpson, T. L., Bihrle, A. M., Potter, H. H. & Gardner, H. (1990). Appreciation of metaphorical alternative word meanings by left and right brain-damaged patients. *Neuropsychologia, 28,* 375-383.

Canessa, N., Gorini, A., Cappa, S. F., Piattelli-Palmarini, M., Danna, M., Fazio, F., & Perani, D. (2005). The effect of social content on deductive reasoning: an fMRI study. *Hum Brain Mapp, 26*, 30-43.

Cartwright, N. (1999). *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press.

Chaigneau, S. & Barbey, A.K. (2008). Assessing psychological theories of causal meaning and inference. In *Proceedings of the Thirtieth Annual Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Christoff, K. & Gabrieli, J.D.E. (2000) The Frontopolar Cortex and Human Cognition: Evidence for a Rostrocaudal Hierarchical Organization within the Human Prefrontal Cortex. *Psychobiology, 28* (2), 168-186.

Christoff, K., Prabhakaran, V., Dorfman, J., Zhao, Z., Kroger, J.K., Holyoak, K.J. & Gabrieli, J.D.E. (2001). Rostrolateral prefrontal cortex involvement in relational integration during reasoning. *NeuroImage 14,* 1136-1149.

Churchland, P.M. (1999). Eliminative materialism and the propositional attitudes. In W.G. Lycan (Ed.) *Mind and Cognition: An Anthology, 2nd Edition.* Malden, Mass: Blackwell Publishers.

Cooper, J.M. & Hutchinson, D. S. (1997). *Plato, Complete Works,* Hackett Pub. Co.

Corbetta, M., Miezin, F. M., Dobmeyer, S., Shulman, G. L. & Petersen, S. E. (1991). Selective and divided attention during visual discriminations of shape, color, and speed: functional anatomy by positron emission tomography. *Journal of Neuroscience, 11,* 2383–2402.

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

Demb, J.B., Desmond, J.E., Wagner, A.D., Vaidya, C.J., Glover, G.H. & Gabrieli, J.D. (1995). Semantive encoding and retrieval in the left inferior prefrontal cortex: a functional MRI study of task difficulty and process specificity. *Journal of Neuroscience, 15,* 5870-5878.

Dolan, R.J. & Fletcher, P.C. (1997) Dissociating prefrontal and hippocampal function in episodic memory encoding. *Nature, 388,* 582–585.

Durrant, M. (1993). *Aristotle's De Anima in Focus*. New York, NY: Routledge Press. Fonlupt, P. (2003). Perception and judgement of physical causality involve different brain structures. *Cognitive Brain Research, 17,* 248–254.

Fauconnier, G. & Turner, M. (2002).*The Way We Think: Conceptual Blending and the Mind's Hidden Complexities.* New York: Basic Books.

Fodor, J.A. (1983). *The Modularity of Mind: An Essay in Faculty Psychology.* The MIT Press.

Fugelsang, J. & Dunbar, K., (2004*).* A cognitive neuroscience framework for understanding causal reasoning and the law. *Philosophical Transactions of The Royal Society of London.Series* B, 359 1749-1754.

Gallese, V. & Metzinger, T. (2003). Motor Ontology: The Representational Reality of Goals, Actions, and Selves, *Philosophical Psychology* 13, 365-388. Goel, V. (2005). Cognitive Neuroscience of Deductive Reasoning. In *Cambridge Handbook of Thinking and Reasoning*, Eds. K. Holyoak and R. Morrison. Cambridge University Press.

Gentner, D., & Colhoun, J. (in press). Analogical processes in human thinking and learning. In A. von Müller & E. Pöppel (Series Eds.) & B. Glatzeder, V. Goel, & A. von Müller (Vol. Eds.), On Thinking: Vol. 2. Towards a Theory of Thinking. Springer-Verlag Berlin Heidelberg.

Gibson, J. J. (1966) *The Senses Considered as Perceptual Systems.* Houghton Mifflin, Boston.

Goel, V. & Dolan, R.J. (2000). Anatomical segregation of component processes in an inductive inference task. *Journal of Cognitive Neuroscience, 12,* 1-10.

Goel, V., & Dolan, R. J. (2001). Functional Neuroanatomy of Three-Term Relational Reasoning. *Neuropsychologia, 39*, 901-909.

Goel, V. & Dolan, R.J. (2003). Explaining modulation of reasoning by belief. *Cognition, 87,* B11-22.

Goel, V., & Dolan, R. J. (2004). Differential involvement of left prefrontal cortex in inductive and deductive reasoning. *Cognition, 93*, B109-121.

Goel, V., Buchel, C., Frith, C. & Dolan, R.J. (2000). Dissociation of mechanisms underlying syllogistic reasoning. *Neuroimage, 12,* 504–14

Goel, V., Gold, B., Kapur, S. & Houle, S. (1997). The seats of reason: A localization study of deductive and inductive reasoning using PET (O15) Blood Flow Technique. *NeuroReport, 8,* 1305-1310.

Goel, V., Gold, B., Kapur, S. & Houle, S. (1998). Neuroanatomical correlates of human reasoning. *Journal of Cognitive Neuroscience, 10,* 293-302.

Goldman, A. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading,* Oxford University Press.

Goldvarg, E. & Johnson-Laird, P.N. (2001). Naïve causality: a mental model theory of causal meaning and reasoning. *Cognitive Science*, 25, 565-610.

Gopnik, A. & Glymour, C. (2002). Causal maps and Bayes nets: A cognitive and computational account of theory-formation. In P. Carruthers, S. Stich, M. Siegal, (Eds.) *The cognitive basis of science.* Cambridge: Cambridge University Press.

Griffin, D.W. & Ross, L. (1991). Subjective construal, social inference, and human misunderstanding. *Advances in Experimental Social Psychology, 24,* 319-359.

Haxby, J. V., Grady, C. L., Horwitz, B., Ungerleider, L. G., Mishkin, M. & Carson, R. E. (1991). Dissociation of object and spatial visual processing pathways in human extrastriate cortex. *Proceedings of the National Academy of Sciences, 88,* 1621–5.

Hegarty, M. (2004). Mechanical reasoning as mental simulation. *Trends in Cognitive Science*, *8*, 280-285.

Houde, O., Zago, L., Mellet, E., Moutier, S., Pineau, A., Mazoyer, B., & Tzourio-Mazoyer, N. (2000). Shifting from the perceptual brain to the logical brain: the neural impact of cognitive inhibition training. *J Cogn Neurosci, 12*, 721-728.

Iacoboni, M. (2008). *Mirroring People: The New Science of How We Connect with Others.* Farrar, Straus, and Giroux.

Iacoboni, M., Woods, R.P., Brass, M., Bekkering, H., Mazziotta, J.C. & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science,* 286, 2526-2528.

Johnson-Laird, P.N. (1998). Imagery, visualization, and thinking. In J. Hochberg (Ed.), *Perception and cognition at century's end* (pp. 441–467). San Diego, CA: Academic Press.

Johnson-Laird, P.N. (1995). Mental models, deductive reasoning, and the brain. In M. S. Gazzaniga (Ed.) *The Cognitive Neurosciences,* MIT Press, Cambridge, MA, pp. 999-1008.

Johnson-Laird, P.N. (1983). *Mental Models: towards a cognitive science of language, inference, and consciousness,* Cambridge University Press, Cambridge. Johnson-Laird, P.N. & Goldvarg-Steingold, E. (2007) Models of cause and effect. In Schaeken, W., Vandierendonck, A., Schroyens, W., and d'Ydewalle, G. (Eds.) *The Mental Models Theory of Reasoning: Refinement and Extensions.* Mahwah, N.J.: Erlbaum. pp. 167-189.

Knauff, M., Fangmeier, T., Ruff, C. C., & Johnson-Laird, P. N. (2003). Reasoning, models, and images: behavioral measures and cortical activity. *J Cogn Neurosci, 15*, 559-573.

Knauff, M., Mulack, T., Kassubek, J., Salih, H.R. & Greenlee, M. W. (2002). Spatial imagery in deductive reasoning: A functional MRI study. *Cognitive Brain Research, 13,* 203-212.

Kosslyn, S.M., Koenig, O., Cave, C. B., Tabg, J. & Gabrieli, J.D.E. (1989). Evidence for two types of spatial representations: Hemispheric specialization for categorical and coordinate relations. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 723-735.

Kroger, J.K., Nystrom, L.E., Cohen, J.D. & Johnson-Laird, P.N. (2008). Distinct neural substrates for deductive and mathematical processing. *Brain Research*, 1243, 86-103.

Kurby, C.A., Zacks, J.M. & Xia, J. (2008). fMRI evidence for the activation of modality-specific images during silent reading. *18th Annual Conference of the Society for Text and Discourse*, Memphis, Tennessee.

Lagnado, D.A. & Channon, S. (2008). Judgments of Cause and Blame: The influence of Intentionality and Foreseeability. *Cognition* 108, 754-770.

Michotte, A. (1963). *The perception of causality* (T. R. Miles & E. Miles, Trans.). New York: Basic Books. (Original work published 1946).

Monti, M., Osherson, D., Martinez, M. & Parsons, L. (2007). Functional neuroanatomy of deductive inference: a language-independent distributed network. *Neuroimage,* 37, 1005-1016.

Noveck, I. A., Goel, V., & Smith, K. W. (2004). The neural basis of conditional reasoning with arbitrary content. *Cortex, 40*, 613-622.

Osherson, D., Perani, D., Cappa, S., Schnur, T., Grassi, F. & Fazio, F. (1998). Distinct brain loci in deductive versus probabilistic reasoning. *Neuropsychologia, 36,* 369-76.

Parsons, L.M. & Osherson, D.N. (2001). New evidence for distinct right and left brain systems for deductive versus probabilistic reasoning. *Cerebral Cortex 11,* 954-65, 2001.

Palmer, S.E. (1999) *Vision science: Photons to phenomenology.* Cambridge, MA: MIT Press Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.

Rizzolatti, G. & Sinigaglia, C. (2008). *Mirrors in the Brain: How Our Minds Share Actions, Emotions, and Experience,* Oxford University Press.

Rorty, R. (1970). In defense of eliminative materialism. In D.M. Rosenthal (Ed.) *The Review of Metaphysics XXIV*.

Roger, M.E., Fugelsang, J.A., Dunbar, K.N., Corballis, P.M. & Gazzaniga, M. (2005). Dissociating processes supporting causal perception and causal inference in the brain. *Neuropsychology, 19,* 591-602.

Rozenblit, L.R. and Keil, F.C. (2002). The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science, 26,* 521-562.

Prado, J., & Noveck, I. A. (2007). Overcoming perceptual features in logical reasoning: a parametric functional magnetic resonance imaging study. *J Cogn Neurosci, 19*, 642-657.

Schacter, D.L., Addis, D.R. & Buckner, R.L. (2007). Remembering the past to imagine the future: The prospective brain. *Nature Reviews Neuroscience,* 8, 657-661.

Schachter, D.L. & Wagner, A.D. (1999), “Remembrance of Things Past”, *Science* 285: 1503–1504.

Shallice, T. & Evans, M. (1978). The involvement of the frontal lobes in cognitive estimation. *Cortex, 14,* 294-303.

Sloman, S.A. (2005). Causal models: How we think about the world and its alternatives. New York: Oxford University Press.

Sloman, S.A., Barbey, A.K. & Hotaling, J. (2009). A causal model theory of the meaning of "cause," "enable," and "prevent." *Cognitive Science*, 33, 21-50.

Sloman, S.A. & Lagnado, D. (2005). The problem of induction. In R. Morrison and K. Holyoak (Eds.). *Cambridge Handbook of Thinking & Reasoning,* New York: Cambridge University Press, pp. 95-116.

Smith M.L. & Milner, B. (1988) Estimation of frequency of occurrence of abstract designs after frontal or temporal lobectomy. *Neuropsychologia, 26,* 297–306.

Spirtes, P., Glymour, C. & Scheines, R. (1993). *Causation, prediction, and search*. New York: Springer-Verlag.

Stenning, K. (2002). *Seeing Reason: Image and Language in Learning to Think.* Oxford: Oxford University Press.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, 12, 49-100.

Walsh, C. & Sloman, S. A. (submitted). The meaning of cause and prevent: the role of causal mechanism.

Whitaker, H., Savary, F., Markovits, H. & Grou, C. (1991). Inference deficits after brain damage. *INS Meeting,* San Antonio.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, 136, 82-111.

Wolff, P., Barbey, A.K. & Hausknecht, M. (2010). For want of a nail: How absences cause events. *Journal of Experimental Psychology: General*, 139: 191, 221.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

Zeki, S. (1993). *A vision of the brain.* Cambridge, MA: Blackwell Scientific Publications. Inc.

Figure Legends

Figure 1. Configurations of forces associated with CAUSE, HELP/ENABLE/ALLOW, and PREVENT; **A** = the affector force; **P** = the patient force; **R** = the resultant force; **E** = endstate vector, which is a position vector, not a force.

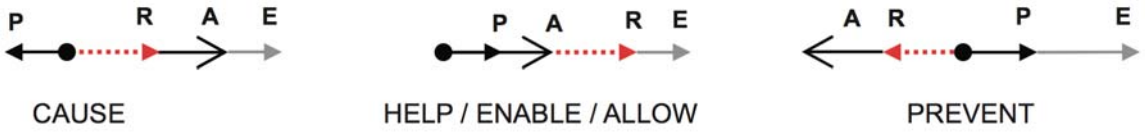


Figure 2. The affector force in the conclusion, **A**, is the affector force in the first relation, **A**. The endstate in the conclusion is the endstate vector from the previous premise. The patient force in the conclusion, **C**, is based on the vector addition of the patient forces **B** and **C** in the premises.

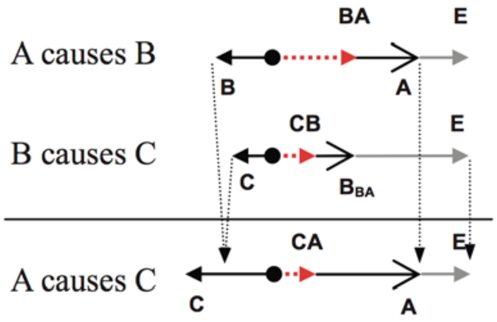


Figure 3. Summary of the main neural predictions of the Force Theory.

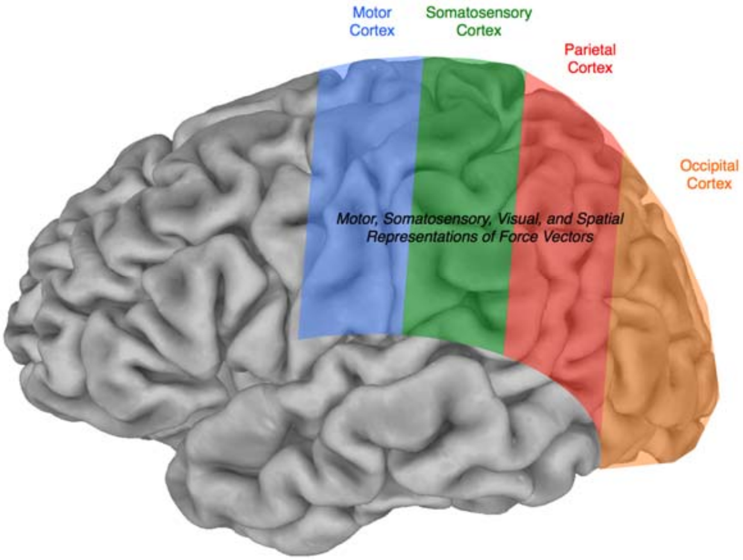

Motor Cortex
Somatosensory Cortex
Parietal Cortex
Occipital Cortex
Motor, Somatosensory, Visual, and Spatial Representations of Force Vectors

Figure 4. Causal Model Theory

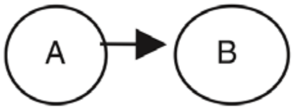


Figure 5. A causal Bayes nets theory of the meaning of *CAUSE*, *ENABLE*, and *PREVENT*.

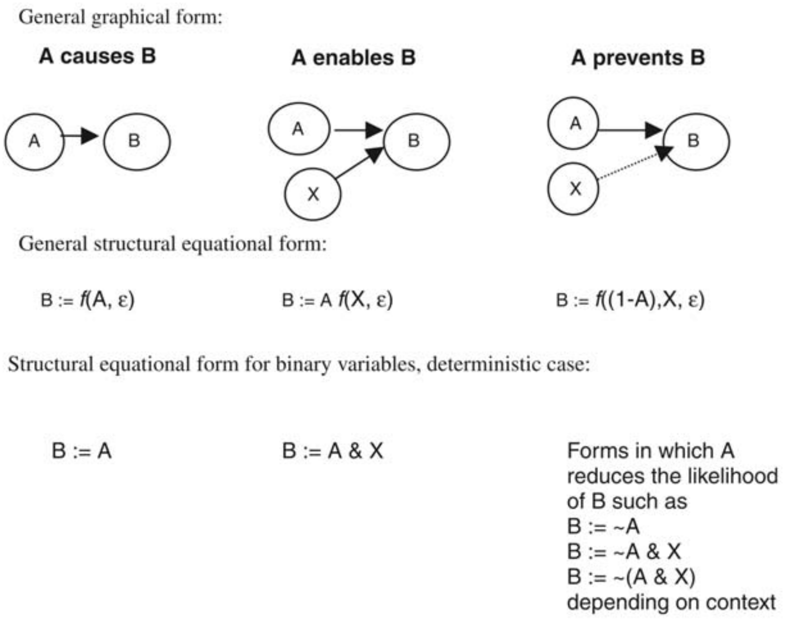


Figure 6. Summary of the main neural predictions of Causal Model Theory.

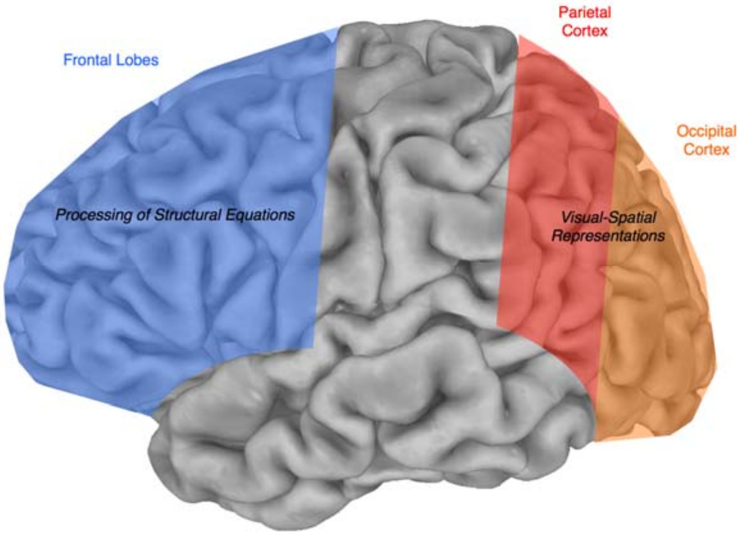

Frontal Lobes
Parietal Cortex
Occipital Cortex
Processing of Structural Equations
Visual-Spatial Representations

Figure 7. Summary of the main neural predictions of Mental Model Theory.

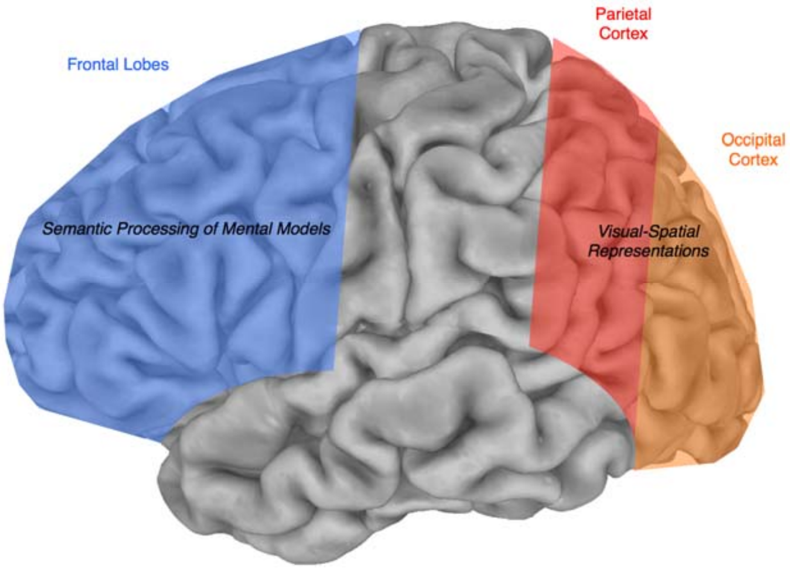
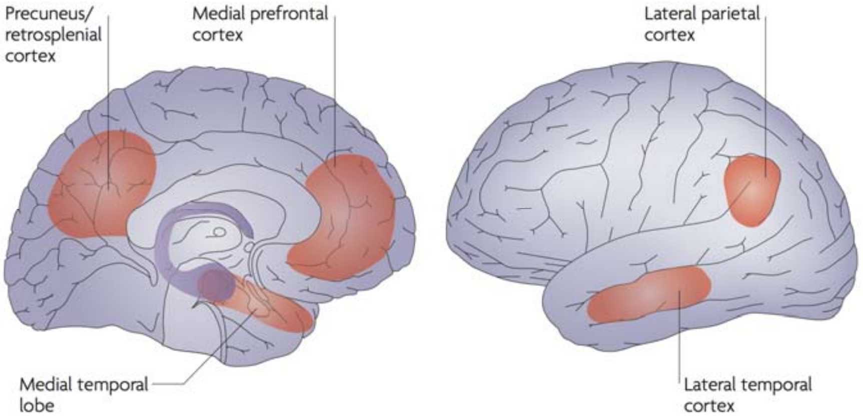
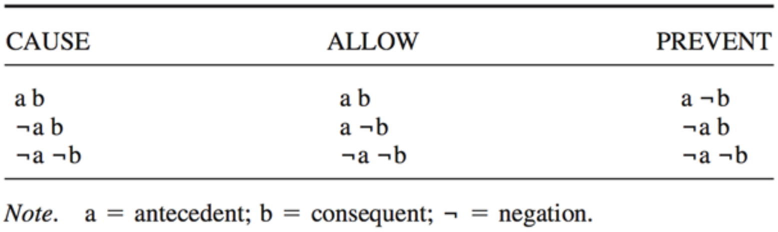

Frontal Lobes
Parietal Cortex
Occipital Cortex
Semantic Processing of Mental Models
Visual-Spatial Representations

Figure 8. Core brain regions for constructing causal simulations (Reprinted with permission from Schacter et al., 2007).



Tables

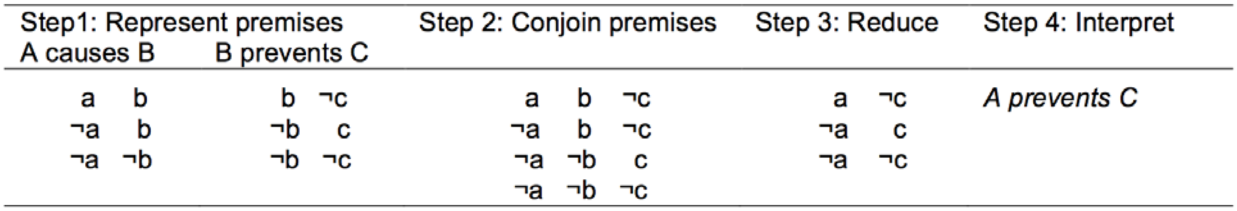
Table 1. Summary of mental models underlying *CAUSE*, *ALLOW*, and *PREVENT*.



CAUSE	ALLOW	PREVENT
a b	a b	a ¬b
¬a b	a ¬b	¬a b
¬a ¬b	¬a ¬b	¬a ¬b

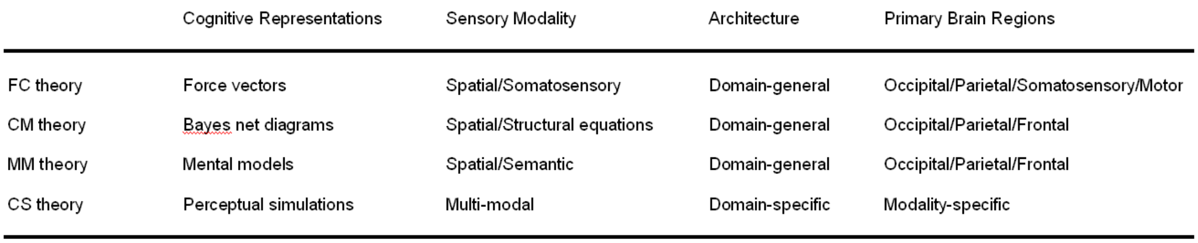
Note. a = antecedent; b = consequent; ¬ = negation.

Table 2. An example of causal reasoning according to mental model theory.



Step1: Represent premises		Step 2: Conjoin premises	Step 3: Reduce	Step 4: Interpret
A causes B	B prevents C			
a b	b ¬c	a b ¬c	a ¬c	*A prevents C*
¬a b	¬b c	¬a b ¬c	¬a c	
¬a ¬b	¬b ¬c	¬a ¬b c	¬a ¬c	
		¬a ¬b ¬c		

Table 3. Neural predictions of the reviewed theories of causal reasoning.



	Cognitive Representations	Sensory Modality	Architecture	Primary Brain Regions
FC theory	Force vectors	Spatial/Somatosensory	Domain-general	Occipital/Parietal/Somatosensory/Motor
CM theory	Bayes net diagrams	Spatial/Structural equations	Domain-general	Occipital/Parietal/Frontal
MM theory	Mental models	Spatial/Semantic	Domain-general	Occipital/Parietal/Frontal
CS theory	Perceptual simulations	Multi-modal	Domain-specific	Modality-specific

Table 4. Summary of findings from 19 neuroimaging studies of deductive reasoning and reported regions of activation corresponding most closely to the main effect of reasoning. Numbers denote Brodmann Areas; RH = Right Hemisphere; LH = Left Hemisphere; Hi = Hippocampus; PSMA =Pre-Sensory-Motor Area Blank cells indicate absence of activation in region. "Stimuli modality" refers to the form and manner of presentation of the stimuli. Cerebellum activations are not noted in the table. Reproduced from Goel (2007).

Studies (Organized by Tasks)	Scanning Method	Stimuli Modality	Occipital Lobes		Parietal Lobes		Temporal Lobes		Basal Ganglia		Cingulate		Frontal Lobes	
			RH	LH	RH	LH	RH	LH	RH	LH	RH	LH	RH	LH
Transitivity (Explicit)														
Goel et al. (1998)	PET	visual, linguistic		19				37		yes		24, 32		45, 46
Goel & Dolan (2001)	fMRI	visual, linguistic	17, 18, 19	19	7, 40	7, 40			yes	yes			6	6, 9
Knauff et al. (2003)	fMRI	auditory, linguistic			7	7	21	21, 38					6	46, 47
Goel et al. (2004)	fMRI	visual, linguistic	18, 19	18,19	7, 40	7	21, 22, Hi	21, 22, Hi					11, 47	6, 9, 46,11
Fangmeier et al. (2006)	fMRI	visual, nonlinguistic			7	40						32	6	6, 9
Transitivity (Implicit)														
Acuna et al. (2002)	fMRI	visual, nonlinguistic			7, 39, 40	39, 40				yes			6, 8, 9, 46	6, 8, 9, 46
Heckers et al. (2004)	fMRI	visual, nonlinguistic			40	40	37, Hi	37, 21		yes		24	PSMA, 6	6, 47
Categorical Syllogisms														
Goel et al. (1998)	PET	visual, linguistic						21, 22				24, 32		45, 46, 47
Osherson et al. (1998)	PET	visual, linguistic	18						yes					6
Goel et al. (2000)	fMRI	visual, linguistic	18, 19	18		7	21/22		yes	yes			45	44, 45
Goel & Dolan (2003)	fMRI	visual, linguistic	17, 18	17, 18				21, 22, 38	yes				6	6, 44
Goel & Dolan (2004)	fMRI	visual, linguistic	18	18, 19	7	37		39	yes	yes			6	6, 44, 45
Conditionals (Simple)														
Noveck et al. (2004)	fMRI	visual, linguistic		19		7		37				32		6, 47
Prado & Noveck (2006)	fMRI	visual, linguistic	18	17	39, 40	40							6, 45, 46	9, 46
Conditionals (Complex)														
Houde et al. (2000)*	PET	visual, nonlinguistic												
Parsons et al. (2002)	PET	visual, linguistic	18	18			21, 37, 39		yes	yes	24	31	10, 44, 9	
Canessa et al. (2005)	fMRI	visual, linguistic	19	19	7, 39, 40	7, 39, 40				yes	32	32	6, 8, 9, 10, 46	6, 8, 9, 46
Mixed Stimuli														
Goel et al. (1997)	PET	visual, linguistic												
Knauff et al. (2001)	fMRI	auditory, linguistic	19	19	7, 40	7, 14	21, 22	21, 22			32	32	6, 9	6, 9

Notes: *Brodmann Areas not provided by authors.