International Journal of Comparative Psychology, 2011, 24, 365-388.

Do Associations Explain Mental Models of Cause?

Itxaso Barberia
University of Deusto, Spain

Irina Baetu
University of Adelaide, Australia

Robin A. Murphy
University of Oxford, United Kingdom

A. G. Baker
McGill University, Canada

The propositional or rationalist Bayesian approach to learning is contrasted with an interpretation of causal learning in associative terms. A review of the development of the use of rational causal models in the psychology of learning is discussed concluding with the presentation of three areas of research related to cause-effect learning. We explain how rational context choices, a selective association effect (i.e., blocking of inhibition) as well as causal structure can all emerge from processes that can be modeled using elements of standard associative theory. We present the auto-associator (e.g., Baetu & Baker, 2009) as one such simple account of causal structure.

Newton's (1687/1934) laws of motion and gravitation describe a world that is deterministic and rule-based. Apples fall with a regularity and obedience to Newton's law that would be the envy of any human law-maker. Post-Newtonian physics is still rule-based, although the rules are now stochastic. Newton deduced his laws but all living organisms have evolved mechanisms that internalize these rules. Learning is a mechanism that allows internal adjustments to the physical rules of the environment. Consequently an animal's behavior can be described by a series of rules that reflect laws of the external physical world or at least transformations or approximations of them.

Rules by their nature have a structure and, as such, adaptations to the rules in the world will appear on the surface to involve propositional or inferential processes even when no such processes are involved. Take a simple classical conditioning experiment from Pavlov (1927), where a metronome is regularly followed by food powder. The rules in the world are: if metronome then food, if no metronome then no food. The well-trained dog will come to salivate in the presence of the metronome. In addition to the world's rule about food, the dog has additional rules and inferences it must make. It must represent some version of the

This work was supported by a predoctoral fellowship awarded to Itxaso Barberia by the Generalitat de Catalunya (with the support of the Comissionat per a Universitats i Recerca del Departament d'Innovació, Universitats i Empresa de la Generalitat de Catalunya and the Fons Social Europeu), a postgraduate fellowship awarded to Irina Baetu by the Natural Sciences and Engineering Research Council of Canada (NSERC), a post-doctoral fellowship awarded to Irina Baetu by the Fonds Québécois de la Recherche sur la Nature et les Technologies, an NSERC Discovery Grant awarded to Andy Baker, and a grant from the Spanish Ministerio de Educación y Ciencia (SEJ2007 – 67409 – C02 – 01). Correspondence should be addressed to A. G. Baker, Department of Psychology, McGill University, 1205 Dr. Penfield Ave., Montreal, QC (andy.baker@mcgill.ca).

following rule: If I eat food then, in the interests of better digestion (homeostasis), I had better salivate; however, because the food regularly follows the metronome, I should prepare for the food by salivating during the metronome and refrain from salivating in the absence of the metronome.

Ethologists might call this description of these rules a functional analysis of the behavioral system but, for our purposes, one only need consider that it describes an adapted sequence of physical rules in the world, although some rules are outside and others are inside the animal. For psychology, there are three major issues that need to be addressed. First, which rules in the world does the animal internalize and how closely does its behavior map onto these rules? Second, how and at what level of abstraction does the animal internalize these rules? Third, how are these rules instantiated in the biology of the animal? These three questions also reflect three levels of analysis of behavior. There has been a long friction in animal learning between these levels of analysis, both concerning which is most important and what should be the content of the levels (e.g., Murphy, Mondragon, & Murphy, 2008). A similar friction exists in other areas of study. In vision, Marr (1982) identified what he called the computational, algorithmic and implementation levels of analysis. The computational level defines the rules that describe the organism's response to the events in the world. For example, it might report a visual illusion and this might be a ruled based phenomenon. The algorithmic level of analysis for instance might involve arguing that form vision is a consequence of Fourier Transformations of the visual world (Cornsweet, 1970). Finally, the implementation describes the neural basis for these abilities, for this paper, we are only interested in the first two levels and will leave the physiological implementation to one side.

Some learning psychologists have even asked if it is legitimate to go beyond the first level of analysis. Radical behaviorists, led by Skinner, questioned whether "theories" or algorithms were useful. In Skinner's, often maligned, "black box" approach, psychologists were entreated to only consider the inputs (stimuli) and the outputs of the system (responses) and the mathematical rules that link them. Both stimuli and responses could be very broadly defined. Although this approach was often, perhaps unfairly, criticized (e.g., Chomsky, 1959; Fodor & Piattelli-Palmarini, 2010), it has stood the test of time. Herrnstein's (1961) matching law is an example of a modern instantiation of this approach and it has considerable generality (e.g., Koehler & James, 2009). An animal's choice of two (or more) alternatives is determined by the proportion or ratio of economic returns connected to the alternatives. Clearly this is a description about rules or contingencies in the world (see also Murphy et al., 2008). Two or more inputs determine the output of the system. Like all "rules," this can be written in propositional and even inferential form, even if some consequences might not be strictly rational. Radical behaviorists were interested in what the animal computed and not in the internal mechanism or algorithm from which this computation might emerge.

Some Background: The North American Behaviorists

In contrast to radical behaviorism a number of other traditions in animal learning did accept the second, algorithmic, level. They also foreshadow the comparison between the associative and propositional classes of explanations of learning that we will discuss further on. Hull, Spence, and Guthrie are representative of one camp and Tolman the other. Hull was a stimulus-response (S-R) psychologist. He and others believed that goal directed behavior could be understood by claiming that animals formed associations between stimuli and responses. These associations were fueled by the temporal order, timing and motivational significance of the stimuli. S-R links could be strengthened by rewards (reinforcement). In modern language, Hull believed that the goal-directed nature of behavior, and hence expectations, were an emergent property of these associative processes.

Goal directed behavior is interesting because it involves a quest for a thing that is not present in the here and now. It is based on an expectation of the future. And a materialist theory involving S-R associations and reinforcement has no direct representation of expectations of future events (see Dickinson & Balleine, 1994). This begs the question: What causes the first step down a maze and fills the gap between it and the expected reward? To solve this conundrum, Hull (1943) called upon the processes of secondary reinforcement and S-R associations. An animal learned to run down a maze to get food through the process of secondary reinforcement strengthening chains of S-R associations between stimuli (including those internal to the animal) along the maze and, rather molecular, responses, presumably steps down the maze, that have been associated with them. When an animal first learns to run down the maze, entering the goal box is rewarding because food is immediately available and this reward both strengthens the tendency to approach the reward but also is paired with stimuli in the maze that are present when the reward is encountered, thus giving them value. These stimuli themselves become rewarding so "acquiring" them rewards approach responses to them. They have a dual role because they also elicit responses through S-R associations. This process is mediated by Hull's anticipatory goal responses and goal stimuli (r_g and s_g). These are an amalgam of the initially consummatory responses and their feedback. These intervening variables provide a formal structure to mediate the expectations in the maze and help generalize initial goal directed behavior throughout the maze and distinguish it from Skinner's simple chaining account. Thus an animal that has learned to run down a maze for food takes its first step down the maze not because it is thinking about future food, but because the first stimulus in the maze generates a step (through an S-R bond) that takes it to the second stimulus that itself, through reinforcement, strengthens the initial bond and so on. The rule in the world that describes goal directed behavior – *There is food at the end of the runway so I had better run down there and get it!* – has emerged from a simple associative process. The Grice box experiment was designed to test this hypothesis (Hughes, Davis, & Grice, 1960). Its rationale was to determine if animals could tolerate a delay of reinforcement for choice behavior when the chain of responses and stimuli was broken or made ambiguous. And, of course, they could not. Goal directed choice behavior could only develop over very

brief delays. Longer delays were mediated by these immediate reinforcement mechanisms rather than an appreciation or expectation of the delayed reward.

Tolman (1932, 1948) took another tack and argued that expectations were not something that needed to be explained, rather they were fundamental primitives of his explanations of goal directed behavior. That is to say, the rule in the world called goal directed behavior was internalized within the animal and was itself an explanation of behavior rather than a behavior that need be explained. To demonstrate the failings of the S-R explanation of goal directed behavior, he devised several experiments directly testing the notion that expectation was a direct consequence of secondary reinforcement and immediate S-R associations. His blocked path experiment illustrates his approach (Tolman & Honzik, 1930). The experiment was done in a maze that had three separate paths to the goal each longer than the other. A schematic diagram of this maze is shown in Figure 1. Once animals were trained in the maze and had experience with all three arms, he blocked either the shortest path to the goal (Block A in Fig. 1) or the two shortest at a common point (Block B). The rationale rule system in the world for this maze was – *If there are no blocks then I should choose the shortest path; If I discover the shortest path is blocked then I should choose the medium length path; If I discover the two shortest paths are blocked then I should switch immediately to the longest path*. Interestingly, the s_g-r_g mechanism has little trouble obeying or generating the first two rules. Because the string of secondary reinforcers and S-R associations grows longer as the path becomes longer, the strength of the association that causes the animal to take its first step down an arm, and hence choose an alternative, is stronger for the shorter arms. So the animal originally (and rationally in the world) chooses the shortest arm. Once he abandons that arm, presumably through extinction, he will be drawn down the second longest arm. Thus the first two parts of the rules for goal directed behavior have emerged from S-R theory without recourse to an expectations or thoughts about future events. At this level the theory has provided an explanation of expectancy. However, S-R theory also predicts that the animal should choose the second longest arm, regardless of whether the shortest or the two shortest arms are blocked, because only the s_g-r_g associations in the chosen short path are extinguished. It is well known that Tolman's rats behaved as if they had an appreciation of the overall structure of the maze because those for whom both shorter arms were blocked were more likely to directly switch from the shortest to the longer, unblocked, path.

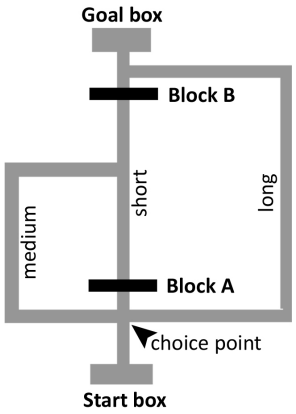

Figure 1. Diagram of the blocked path maze used by Tolman and Honzig (1930).

This experiment is relevant to our considerations because it reveals several enduring features of the analysis of cognitive learning systems. First, it shows how quite complex and "rational" behavior can emerge from a system that does not have cognitive representations of goal directed behavior. S-R psychology does not have expectations or an appreciation of the overall structure of the environment or a particularly complex planning system built in. It is parsimonious and these interesting features emerge from it. This very parsimonious system is "correct" on two of its three choices. Second, the alternative position to simple associationism is often couched in propositional or inferential terms. When this is done it is not clear if it is being claimed that there is a propositional machine or actor inside the organism that is doing this rule-based reasoning. That is, is this a propositional algorithm? Or is it just a description of how the animal's behavior maps onto certain rules of the world? If so an algorithm for how this is done must be developed. Tolman held the former position but it is not always so clear with others. Third, it is hard to disconfirm these propositional theories, because, like many such systems, they depend on the premises. If an animal fails to switch immediately to the longest path this may be because S-R theory is correct, and it is being dragged around the world by fractional anticipatory goal stimuli. Equally this might happen because the rat has not developed an appropriate appreciation of the maze (called a cognitive map by Tolman, 1948), and thus simply made a mistake. Depending on the operator's assumption, a propositional system can either act rationally or not based on the premises, but this in no way disconfirms the propositional mechanism, unless, of course, the theorist clearly specifies and fixes the underlying premises (see also Mitchell, De Houwer, & Lovibond, 2009). In these experiments this is difficult to do and thence often leads to circularity. The behavior confirms the premises and the propositional account. If the animal fails, then the premises may be different and thus the propositional account is still confirmed. Following the work we have just described came the cognitive revolution that rejected many of the tenets of traditional North American behaviorism or associationism.

Brewer and the Cognitive Revolution

One consequence of the cognitive revolution was to focus conditioning research on the representational questions that the algorithmic level analysis might answer. At least partly, this approach was inspired by readers of Chomsky's (1959) review of Skinner's Verbal Behavior (1957), which argued that the traditional associative approach was impoverished and even circular (see also Fodor & Piattelli-Palmarini, 2010). For instance, Brewer (1974) argued that conditioning is not a mechanistic bottom up process driven directly by associations and reinforcement; rather, in humans at least, it is always cognitively penetrable. While it is easy to quibble with his characterization of associationism, the important point is that there is a great deal of research supporting his position. There is a strong *prima fascia* case for his position. In simple language, the notion of cognitive penetrability is that all conditioning represents an internalization of the rules of the conditioning experiment. For example, if a person learns an eyeblink response, she does so because she learns the CS-US rule. And if asked an appropriate question she can usually report the rule and behave as if she has internalized it as a proposition and make inferences from it. There is a great deal of research on this point, but the general theme can be understood with two examples. Brewer pointed out that, in many experiments, only some of the people develop a conditioned response. Interestingly, those people who have developed the response are much more likely to be able to report the rule (e.g., tone is followed by shock and light not) than those who have not. This is consistent with the claim that conditioning occurred because they internalized the propositional rule and that that rule is available to them. A second finding is also interesting. It seems that a well-conditioned participant, just like a rat, can have her responding extinguished. However, extinction can often be established by simply informing the person that the shock will no longer be delivered. Disconnecting the electrodes is even more effective. This again implies that the behavior is driven at the time of the test by a propositional inference – *If tone then shock, so I should blink.* However the premise is changed by informing the participant of the absence of shock leading to a new more cheerful inference.

There are of course many possible objections to the notion of cognitive penetrability. Cognitive representations, including awareness, are supposed to be an emergent property of associations, so it would not be surprising that only those who behaved (i.e., formed associations), are also aware. Moreover, telling people that the shock will no longer occur may engender generalization from other experience and certainly changes in the context (Bouton, 2004), both of which can immediately change behavior. But, crucially, there is a large corpus of behavior supporting this propositional framework and, as we will discuss, the propositional position has been very effective in generating empirical results that were not obviously coming from within the associationist framework.

The enthusiasm for the rejection of associationism waned with the development of connectionist and similar models, although these were rarely identified with their behaviorist ancestors. Nonetheless, the idea of cognitive penetrability has been kept alive both in animal research and human research (De Houwer, 2009; De Houwer, Beckers, & Vandorpe, 2005; Lovibond & Shanks,

2002; Mitchell et al., 2009). Indeed, it has even been argued that, given existing positive research in humans and rats, the most parsimonious position is to claim that all conditioning in all organisms is propositional in nature (De Houwer, 2009; Mitchell et al., 2009). Although it must be acknowledged that the authors wonder at the plausibility of this claim. From our discussion above, this is not an issue if they are talking at the computational level. If so, it is easy to accept the argument that a planarian might compute inferences (perhaps in some associative way) with its simple nervous system. This level of computation might be a little simpler than a human's. It is not so clear how this might be a meaningful algorithm.

In the subsequent sections of this paper we will discuss three lines of our research that we believe will help to reconcile the propositional accounts of learning with associative explanations. To foreshadow our eventual conclusions, we will argue that propositional accounts can offer useful computational accounts of behavior. Moreover they are a very powerful heuristics for generating new experiments. However, they are rarely sufficiently well-specified to be considered algorithms or explanations of behavior. Thus, they are difficult to disconfirm. The potential for disconfirmation is the litmus test for a scientific theory or algorithm. Finally, we will argue that simpler mechanistic algorithms can be developed to explain this propositional behavior. In the first section, that describes interventions to use the most informative context in causal discovery, we will show how a propositional analysis motivated these experiments, but also how some wrinkles in the data illustrate the weaknesses of underspecified propositional accounts. In the second section, investigating blocking of conditioned inhibition in causal reasoning, we will test some predictions of a popular, inferential, propositional theory and find them wanting. Again we will illustrate the difficulty disconfirming various versions of this approach. In the third section, we will discuss some experiments investigating novel predictions about how participants construct causal chains from negative and positive individual causal links. We will then show how these predictions that came from a logical propositional analysis emerge from a simple associative model. The model also generates important philosophical features of causal reasoning such as temporal precedence and timing of cause and effect.

Choosing the Most Informative Context

One of the advantages of having a mental model of the mechanism of a cause, over having a simple representation of its effect in a specific context, is that it will allow the observer to plan informative interventions. Indeed, the use of such strategies, sometimes called causal graph surgery, has been argued to be one of the compelling arguments in favor of mental models of cause and the "Bayesian" movement in causal reasoning (e.g., Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). That is, if people assume that causal power exists within the cause, this implies that they will be aware of the situations in which this power will be more effective. The differential informative value of a set of interventions is especially obvious in the simplest case in which the target cause and its effect are binary, and the influence of the target cause and the sum of the effects of the unknown alternative causes (i.e., the causal context) are independent. In this

situation clear predictions can be made about the expected rational preference for some interventions over others. And this expected preference is different for generative and preventive causes.

A generative cause is analogous to an excitatory stimulus in conditioning; it is one whose presence signals an increase in the likelihood of an outcome. The actions of a generative cause will be masked in the presence of other alternative causes of the same outcome. Consequently, the best situation to observe a target cause's effectiveness is one with no effective, or at least weakly effective alternative generative causes. Thus, the best interventions would involve choosing to introduce a generative cause in contexts in which the influence of other potential generative causes is weak. For example, if we want to study if a new public service announcement effectively promotes the use of helmets by bicycle riders, the best option would be to test it on a group of people who normally do not wear helmets. At the other extreme, it might be quite uninformative to test its effect on people who already use helmets.

Preventive causes reduce the likelihood of an outcome and thus are analogous to inhibitors in conditioning. For them to show their efficacy there must be some outcomes to prevent. Thus, the best contexts to observe their actions will be ones in which the effect is frequently generated by the context or alternative generative causes, because there will be more instances in which the target cause may show its preventive influence. For example, if we want to study if a drug effectively prevents headaches, then it would be better to give it to people who regularly have headaches. Testing the drug in people who never have headaches would give no information about its effectiveness. In fact, Wu and Cheng (1999) have shown that people tend to report that causal conclusions are not possible in these extreme situations with effect ceilings or floors. We carried out a series of experiments designed to test if people actively choose more informative and avoid less informative interventions when they are allowed to do so (Barberia, Baetu, Sansa, & Baker, 2010).

We studied the way people would intervene in order to choose the most informative context to discover a potential causal relationship. In order to show people the effectiveness of the contexts or alternative causes, we exposed participants to contexts with different outcome base rates. The base rate is simply the probability or likelihood of the outcome in the absence of the target cause. Subsequently, we asked participants to assess the influence of a potential target cause on the effect. To do so, on each trial, the participants could introduce the cause in one of the previously trained contexts and observe whether or not the effect happened. This strategy differs from the traditional causal discovery task in which participants simply observe contingencies, usually with no context switch (e.g., Vallée-Tourangeau, Murphy, Drew, & Baker, 1998). The scenario involved evaluating whether some unknown folk medicines brought back from the Amazon by a group of scientific explorers could provoke, prevent, or have no influence in the probability of strokes. The "medicines" could be generative causes, that is, they could generate strokes, they could be preventive causes, that is, they could prevent strokes, or they could be ineffective. The substances could be tested in several populations of patients that differed in their genetic predisposition to, and hence their base rate of, strokes. Therefore, the folk medicines were the target causes,

having a stroke was the effect and the different populations were the different contexts in which the causes could be introduced.

The participants initially learned about the likelihood that patients of each genetic type would have strokes by observing records of individual patients who had not been exposed to the folk remedies. To maintain their interest, participants predicted if each patient would have a stroke or not. They were then given feedback. In the subsequent phases participants were "given" 60 doses of each substance. They were then permitted to choose to administer each dose to a patient from one of the previously trained genetic types. On each trial, one dose of the substance appeared on the computer screen, together with the picture of a patient from each genetic type. Participants decided which patient would receive the substance by clicking on the patient's picture. They then predicted if the patient would have a stroke after receiving the substance. Once the prediction was made and feedback received, the next trial appeared on the computer screen. Participants' estimates of the causal status of the contexts and the target causes were recorded, but the critical data here were the proportion of observations they chose to make on each population.

The predictions from the causal model or Bayesian perspective were straightforward. Participants should preferentially choose those genetic types in which the empirical medicine-stroke contingency will more closely reflect the influence or power of the medicine. This implies choosing, for a generative harmful medicine, the genetic type showing the lowest base rate of strokes and, in the case of a preventive medicine, the genetic type with the highest probability of suffering strokes.

In one experiment participants were pre-trained with three contexts with different base rates (BR) of the effect: a population of a genetic type that never had strokes (BR = 0), a population that had strokes half of the time (BR = 0.5) and a population always having strokes (BR = 1). After learning about the three base rates, participants were presented with three substances that could potentially cause or prevent the strokes. They were instructed to find out about their real influence. There were three folk medicines. There was a generative medicine that, in the absence of alternative causes, produced strokes half of the time. This medicine had a causal power (Cheng, 1997) of p = 0.5. The, second, preventive medicine prevented the strokes half of the time in a context in which strokes always occurred (p = -0.5. Although Cheng represents all powers as positive, we use a minus sign to identify preventative powers). Finally, there was a neutral medicine that did not influence the probability of the strokes (p = 0). This third substance acted as a control. Figure 2 shows the results of this experiment (Barberia et al., 2010, Experiment 2 - Group Deterministic). As can be observed in Figure 2, for the neutral control substance, there was a preference for the medium base rate context, BR = 0.5, maybe because only in this population the potential increase or decrease in the probability of the effect could be simultaneously observed. Most importantly, and as expected, participants showed a preference for the low base rate (BR = 0) population when testing a generative substance, and a preference for the high base rate (BR = 1) population when testing a preventive substance.

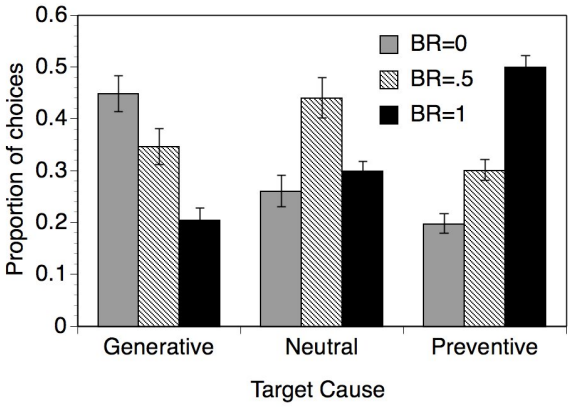

Figure 2. Proportion of choices of the low, medium, and high base rate contexts for the Generative, Neutral, and Preventive target causes, respectively (data from Barberia et al., 2010, Experiment 2 - Group Deterministic).

The results of this experiment seem consistent with the expectations of propositional reasoning. Mental models or "Bayesian" accounts do suggest that people will intervene to choose the most informative context. However, there is a problem. It will be observed that, while the participants preferred the more informative context, they still continued to choose the least informative context a substantial proportion of the time. Indeed, they chose the two less informative contexts about 50% of the time. The propositional account has no simple explanation of this result. However, it might be possible to argue that the participants were using "higher" level strategies such as a belief that the context might interact with the outcomes. They might also believe there is some pattern of context choices that might "explain" the occurrence of all the outcomes. The latter argument is weakened greatly by the fact that elsewhere we have reported other experiments in which we used deterministic causes (Barberia et al., 2010, Experiment 1). With deterministic causes (i.e., p = 1 or -1), the probability of outcomes was either 0 or 100% so there was no pattern, yet people still chose the less, and sometimes the entirely uninformative context, a substantial proportion of the time as they have done in all of our experiments.

However, the point here is to demonstrate both a strength and a weakness of the propositional explanations of behavior. They provide a very useful heuristic for guiding research. It is not clear that we would have chosen to study context choices without such a framework. Nevertheless, they are very difficult to disconfirm. It is the nature of propositional logic that, if one set of premises is untrue, there are many others that are true. Unless serious work is done to constrain the number and class of premises, the propositional accounts approach tautology. Alternatively, they are folk psychology. We now go on to, very briefly, describe some experiments we have done that analyzed a popular propositional description of cue competition or blocking in causal reasoning.

Blocking of Conditioned Inhibition

A number of recent studies asked if people make rational inferences when they observe multiple cues that predict the same outcome. When there are multiple possible causes of an outcome or, in conditioning, multiple predictors (conditional stimuli) of an outcome (unconditional stimulus), cue competition occurs. A common example of this competition is blocking (Kamin, 1969), in which the presence of a stronger or a previously trained predictor reduces judgments or responding to a moderate or new predictor. Propositional inferential accounts argue that this blocking occurs because of the inherent redundancy of the blocking design. Because the blocking stimulus already predicts, or is a cause, of the outcome, then the other stimulus must not be. One can easily see that there are a number of implicit premises about the nature of cause within this framework, but our goal is not to question them here.

The inferential account is much richer and more complex than the simple frame just outlined. Inferential theories argue that people's causal judgments, and their decisions about the possible redundancy of the target cause, are influenced by constraints on the maximum or minimum magnitude of the outcome observed. According to these theories, an observer can most efficiently analyze the influence of an excitatory, or generative, cause only if there is room for the target to actually, and observably, influence, usually increase, the outcome magnitude. For instance, in a blocking design in which a cause, or cue, is followed by an outcome both when presented alone (A - outcome) and when presented with another potential cause, B (AB - outcome; Kamin, 1969; Shanks, 1985) the status of B is ambiguous because it always occurs in the presence of A which has already been established as a cause of the outcome. If this outcome is of maximum observed strength, it would mask the ability of the target to show its effectiveness as a cause because the outcome already has a known effective cause. The effect of B might be disambiguated, however, if the participant could reasonably assume that a compound of two effective causes might generate a stronger effect than either presented alone. Thus, if B is an effective cause, then a larger outcome should be expected when causes A and B are present, if it is assumed that causes have additive effects. Since in a normal blocking procedure the same outcome follows both A and AB, this provides stronger evidence that B is not an effective cause. This strong inference, however, should not be made if a stronger effect than that which occurred on A trials is not possible. Although, as mentioned above, the weaker inferential structure based on simple redundancy of predictors might still operate. Nonetheless, if A alone is followed by the maximum possible outcome, then the effectiveness of B cannot be determined because of a ceiling on the magnitude of the outcome. Consistent with this idea, a number of studies reported stronger blocking effects (i.e., weaker ratings for B) if A and AB were followed by an outcome smaller than the maximum possible outcome, than if the maximum possible outcome occurred on A and AB trials (Beckers, De Houwer, Pineno, & Miller, 2005; De Houwer, Beckers, & Glautier, 2002; Vandorpe, De Houwer, & Beckers, 2005; and, in rats, Beckers, Miller, De Houwer, & Urushihara, 2006).

Simpler versions of associative theories, on the other hand, anticipate no such influence of outcome magnitude and thus do not account for this result. The

simple delta rule used by many associative theories, such as Rescorla and Wagner's (1972) model, changes associations if there is a discrepancy between the actual and the expected outcome, so it operates in the same way regardless of whether there is the possibility of a larger outcome. This learning rule predicts that blocking should occur to the same extent regardless of whether there is a possibility for the outcome to be further increased. Hence, the reported outcome magnitude effects on blocking have been taken as evidence that causal discovery relies on inferential rather than associative processes. Disconfirming this associative prediction has been a major impetus for the inferential blocking experiments we have described.

No one would question whether people actually make inferences or test logical syllogisms when instructed to do so. However, the important question is whether this is the fundamental cognitive structure of all conditioning and of causal reasoning. And it has been argued that it is (De Houwer, 2009; Mitchell et al., 2009). However, many, but not all, of the experiments testing the inferential reasoning account have used very simple experiment designs and possibly leading instructions. When we used very simple instructions and more complex designs, we found a different pattern of results. With minimal instructions and more complex tasks (our participants learned about many cues simultaneously) we found similar, strong, blocking effects regardless of outcome magnitude (Baetu, 2009), as anticipated by most associative theories but not by inferential theories. According to some instantiations of inferential theories, learning that a cue is followed by an outcome requires fewer cognitive resources than making a blocking inference, hence, if the complexity of the task prevents participants from making a blocking inference, then ratings for the target cue B should be high (De Houwer, 2009; Mitchell et al., 2009). Thus, these inferential theories predict that if the task is too complex, blocking should not occur regardless of a ceiling in the outcome level. Our finding of a blocking effect regardless of outcome magnitude was clearly inconsistent with this prediction.

More interestingly, we have extended our findings concerning blocking with generative causes to preventative, or inhibitory, relationships (Baetu & Baker, 2010). We used a blocking of inhibition design analogous to the generative one described previously. A cue was followed by the outcome on its own (A - outcome), but not in the presence of various potentially inhibitory cues (AB - no outcome, ABC - no outcome, ADE - no outcome). B is a traditional conditioned inhibitor because it prevents the outcome caused by A. It is an unambiguous preventive cause in the AB compound. C, on the other hand, is potentially a blocked inhibitory cue because it always co-occurs with B and B already predicts the outcome's omission. D and E are control cues for blocking of C because neither has been trained separately with A as B was, but they would also demonstrate overshadowing of conditioned inhibition, because together they predict the outcome's omission. They are the appropriate control cues to determine whether learning about C is blocked by B: D and E are always trained in compound with another inhibitory cue like C, but, unlike C, neither D nor E is paired with a cue that predicts the outcome's omission on its own.

According to inferential theories, one can evaluate the influence of a preventive cause most efficiently if there is room for it to decrease the outcome

magnitude (Melchers, Wolff, & Lachnit, 2006). Thus, if B already reduces the outcome level to its minimum in the AB compound, there is no further room for C to decrease the outcome magnitude. In this case, C's possible inhibitory strength might be masked by a floor in the outcome level. Conversely, if there is a possibility for the outcome magnitude to decrease below the level of the outcome on AB trials, then one should be certain that C has no causal power or effectiveness when the same outcome level occurs on ABC trials. Thus, C should be more likely to be blocked when there is no floor on the outcome magnitude. Associative theories, on the other hand, predict blocking of cue C regardless of whether there is a possibility for the outcome level to be further decreased. In this experiment we directly compared the level of blocking in a group in which AB brought the outcome level to the "floor" with one in which there was room for C to act in the ABC compound. We used a scenario from Melchers et al. (2006) in which participants discovered whether various food cues influenced a hypothetical hormone level (the outcome) in a patient. As in Melchers et al. (2006), we manipulated the possibility of a lower outcome magnitude by presenting participants in one group only with foods that increased or caused no change in the hormone level (Group Floor), while a second group also observed foods that decreased the hormone level (Group No Floor).

Following training with the above cues and compounds as well as other control cues (Baetu & Baker, 2010), we assessed the inhibitory strengths of the cues of interest using two tests. In the first test we followed the tradition in causal reasoning and asked the participants to directly rate the strengths of the inhibitors. Negative ratings represent preventive causes. This method of assessment is not available in experiments with rats but the second, summation, test is. In the summation test we paired the causes of interest with a novel excitatory cause and observed if the causal strength of the compound was weaker than the strength of the excitor. Would the inhibitory cues inhibit the excitatory strength of the novel excitor? This test is interesting not only because it is directly comparable to tests used with rats, but also because one of the characteristics of causal power is that it is a property of the cause and should transfer to novel situations or contexts. It is this generality of causal power that is the main justification for theories that represent causes as mental models (Waldmann, Hagmayer, & Blaisdell, 2006).

The results of our experiment are clearly consistent with the predictions of associative models. The top panel of Figure 3 shows the first, direct, test for inhibition. This panel shows the ratings given to B (Inhibitory), D and E (Oversh.), C (Blocked), and a neutral, control, cue (Neutral) by two groups of participants, one which experienced a lower outcome level than the one shown on AB and ABC trials (Group No Floor), and one that did not (Group Floor). Both groups showed a similar blocking effect: Ratings of the blocked cue are closer to zero than those of the overshadowed cues. It also shows the predicted overshadowing effect whereby the overshadowed cues were less inhibitory than B. And, again as the simple associative model predicts, the floor manipulation had little effect. It seems the potential ambiguity that arises when the AB compound was on the floor did not interfere with blocking or overshadowing. The summation tests in which the cues were tested in compound with a cue trained to predict the outcome on its own are shown in the lower panel of Figure 3. The results of these tests are similar to the

single-cue ratings: The blocked cue reduced the expected outcome level to a lesser extent than the overshadowed cues in both groups. Again there was evidence of the predicted overshadowing effect. There were, however, overall differences in the ratings given by the two groups since they experienced different outcome levels, which might render between-group comparisons difficult. Nevertheless, there was no reliable difference between the blocked and the neutral cues, demonstrating that C was maximally blocked in the two groups.

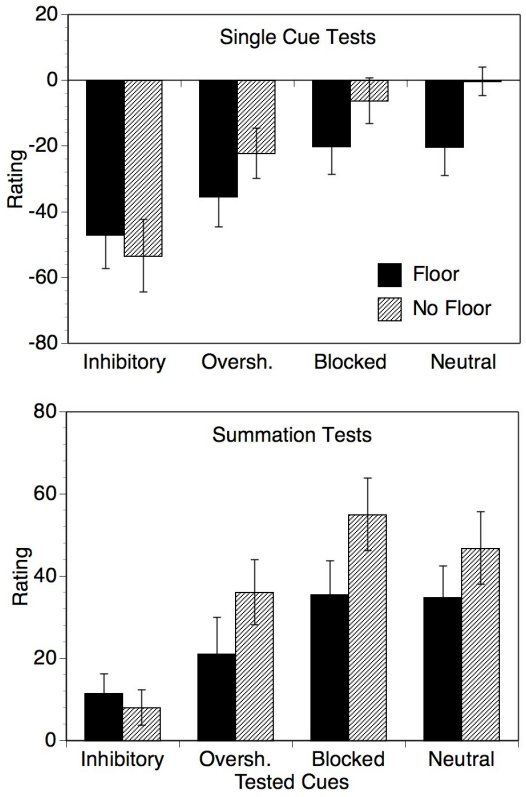

Figure 3. Ratings of the target cues tested individually (upper panel) and in summation tests (lower panel) in Groups Floor and No Floor of Experiment 2 of Baetu and Baker (2010).

Overall, our finding that, in complex learning tasks, blocking of excitatory and inhibitory cues occurs regardless of a ceiling or floor in the outcome magnitude undermines the statement that associative mechanisms play no role in causal learning and that learning effects such as blocking only occur to the extent that inferential reasoning takes place (De Houwer, 2009; Mitchell et al., 2009). Associative-like processes do seem to play a part in causal discovery. What we take from this experiment is that, at the very least, the generality of the predictions of the inferential account is in question. We should also mention that the initial inferential model we described which only makes inferences about the redundancy of the cues and not the absolute level of the outcome could handle our results. However, this demonstrates once more that these accounts are, at the very least, quite difficult to disconfirm. Again propositional models are a good heuristic for

generating research but at their present level of development do not provide good formal models of behavior. We will now discuss some research on learning causal chains (i.e., A lead to B leads to C) and an associative model that predicts this behavior.

Forming Causal Chains Form the Links

One of the critical features of propositional theories of cause is causal precedence. Causes must logically come before effects. Thus, an appreciation of the ordering and timing of events is critical to an understanding of causes. It should not go un-noted that, likewise, it has been known since the days of Pavlov that timing and ordering of events is crucial to conditioning. Similar to causal reasoning, when the Conditional Stimulus (CS) comes after the Unconditional Stimulus (US), little anticipatory excitatory conditioning is observed. We have been developing an associative model that can model the timing and ordering of stimuli. In addition, we have carried out experiments that ask whether participants can form "causal" chains from links that have only been observed in isolation.

The design of our experiments could be summarized with the following syllogism: Participants first learned *If A then B* and *If B then C*; they were then asked if they subsequently inferred: *If A then C*. Instead of being asked to reason about propositions, however, our participants discovered the relationships between A and B and between B and C in a trial-by-trial manner similar to what would be done in a conditioning experiment. That is, they observed instances of the A-B link, intermixed with other instances of the B-C link. Rather than reasoning about chains in which the links were deterministic, the participants were asked to reason about probabilistic links. In these links the two events could occur together or apart. With such an arrangement we could program positive or generative links in which the first event predicted the presence of the second event or we could program inhibitory or preventive links in which the second event was less likely to occur when the first was present. This arrangement was instantiated by a display of three virtual lights on a computer screen. On any trial only two lights were visible and the third was occluded so the participants could not know its state. We did this to maintain the fiction that there was a three light chain and that the participants were only observing two of the three lights on any trial. Following this training, participants evaluated whether A would be followed by C or whether it would prevent C.

The syllogism described above involves a simple chain in which there are positive relationships between A and B and between B and C. We were particularly interested, however, in the way people would reason about chains that include one or two negative links. If A was often followed by B, but B prevented C from happening, would people still expect C to follow A? More interestingly, what would they infer about the A-C relationship if A prevented B from occurring and B prevented C from occurring? From a rational perspective, people should infer that A prevents C in the first case because it enables B to prevent C, whereas they should infer that A causes C to occur in the second case because it prevents B from preventing C. The second event in each link could occur on its own in the absence of the first event, and this implies that some, perhaps hidden, alterative cause of the

second event exists. So in this case, inhibitory B would inhibit C events that were caused by these alternatives, so its elimination would increase the probability of C. In fact, we have shown that this kind of reasoning can emerge from Bayesian principles if one assumes that A acts upon C only through the influence of B (Baetu & Baker, 2009). Cases in which the chain is not made up of only positive links are interesting because they rule out certain confounding explanations. For example, if all the relations are positive then participants may report a positive A-C relationship after having observed positive A-B and B-C links because they have been exposed exclusively to positive relationships. Alternatively, they may simply base their evaluation of the A-C relationship on an average of the two links. Or they may generalize their judgment from any single link in the chain.

Through the mechanism described above (i.e., A prevents B which would otherwise prevent C, so presenting A increases the likelihood of C), reporting a positive A-C relationship after having observed negative A-B and B-C relationships would be rational and it would rule out all of these alternative explanations. Furthermore, participants should report a relationship between A and C only if they perceive nonzero causal links between A and B and between B and C. It only takes one link to break a chain. Participants should be able to detect cases in which one or both links in the chain are "broken," i.e., chains in which, for example, A might influence B, but B would have no influence on C (C would be equally likely in the presence and in the absence of B). In that case, one should rationally infer that A should have no influence on C. We investigated this possibility by having links in which the two events of a link occurred independent of one another, that is to say when they were uncorrelated.

It turns out that people behave rationally in all these cases. In our experiments the A-B and B-C contingencies were positive, negative, or zero, with the constraint that each of the lights turned on on 50% of the trials. For the positive links the conditional probability of the second event in the presence of the first [i.e., P(Event2| Event1)] was 0.8 and the probability in its absence was 0.2 so that the overall contingency (i.e., the difference between these) is $\Delta p = 0.6$ [where Δp = (P(Event2| Event1) – P(Event2 | noEvent1). For the preventive links these probabilities were reversed. Following 40 randomly intermixed A-B trials and B-C trials, participants were asked to evaluate the effect of A on C: whether it would turn C on, prevent C from turning on, or whether it would have no effect on C. Figure 4 shows the mean ratings of the influence of A on C reported on a scale ranging from -10 to +10 (Baetu & Baker, 2009, Experiment 2). When participants observed positive A-B and B-C contingencies (Treatment PP in the figure) or when they observed negative A-B and B-C contingencies (Treatment NN), they inferred that A would turn C on. Conversely, when one of the observed contingencies was positive and the other negative (Treatments PN and NP), they concluded that A would prevent C from turning on. They also inferred that A would have little effect on C if one or both experienced contingencies were zero (Baetu & Baker, 2009, Experiments 1A and 1B; data not shown in the figure).

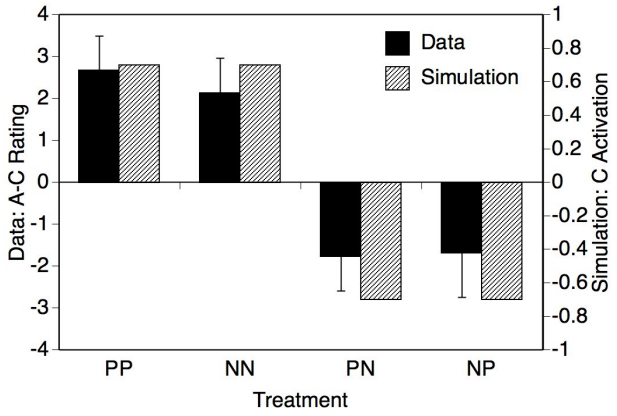

Figure 4. Ratings of the influence of A on C in Experiment 2 of Baetu and Baker (2009) and simulations with the auto-associator.

Our participants behaved rationally and this is consistent with a Bayesian analysis. We were more interested, however, in demonstrating that this kind of rational behavior could emerge as a result of processes that are not constrained by rational premises, but rather from simple associative processes. To do this, we asked whether a simple connectionist model, an auto-associator network implementing a prediction-error learning rule (McClelland & Rumelhart, 1988), would behave in a way similar to our participants and generate a representation of the complete chain from experience with the two individual links of the chain. The network consisted of a single layer of units that might become connected to each other if the stimuli represented by these units co-occur in close temporal proximity. Associations between units allow activation in one unit to spread to others. The appendix briefly describes the way temporal information is represented in the model; the complete model specifications can be found in Baetu and Baker (2009) and McClelland and Rumelhart (1988).

The structure of the network was inspired by the traditional analysis of conditioning experiments. It consisted of six interconnected units. There was one unit representing each of the three events (A, B, and C). There was a unit representing the general context. Finally there were two units representing the context of the A-B trials and B-C trials (Fig. 5). The different trial types were discriminably different because of the presence of the object that would occlude the state of either light A or C depending on whether it was an A-B trial or a B-C trial. Thus, it makes sense to have different trial type contexts. The presence of context units is crucial because, just as has been shown in conditioning inhibition, with negative CS-US correlations (Baker, 1977) the context is critical for modulating contingency learning (Murphy & Baker, 2004; Vallée-Tourangeau et al., 1998).

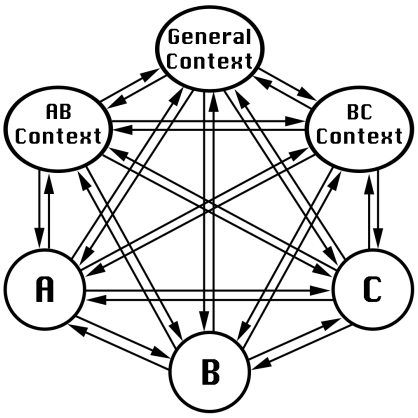

Figure 5. This figure illustrates the structure of the auto-associator composed of six units: three units representing the three lights (A, B and C), a general context unit, a unit representing contextual cues present only on A-B trials, and a unit representing contextual cues present on only B-C trials. The arrows represent all the possible unidirectional connections that might develop between the six units. Figure adapted from Baetu and Baker (2009).

Once this network had been "trained" we asked it about the A-C chain by activating the A unit and monitoring C activation. Figure 4 also shows the results of these "queries" and it can be readily seen that they are consistent with the participants' ratings of the A-C relationships in all treatments. The networks also behaved appropriately when trained with one or two zero links: Unit A failed to activate unit C, which is analogous to our participants reporting that there was no relationship between lights A and C in Experiments 1A and 1B of Baetu and Baker (2009). In addition to carrying out these tests, we investigated the strengths and polarities of the associations in the net. What we found was that the network had "discovered" the causal structure of the events. There were strong and appropriate excitatory and inhibitory connection strengths or associations in the correct direction between the events A, B, and C. The associations involving the context units and the associations in the incorrect direction between A, B and C were much weaker (Fig. 6 shows an example of a trained network). What this means is that in the case of the simple syllogism between positive events described at the beginning of this section when asked "If A?" the network answers "then C!" but if asked "If C?" it does not answer. And this is just what would be expected by propositional theories that posit mental models of cause; but it is done by a simple associative net using standard conditioning assumptions with no formal propositional structure.

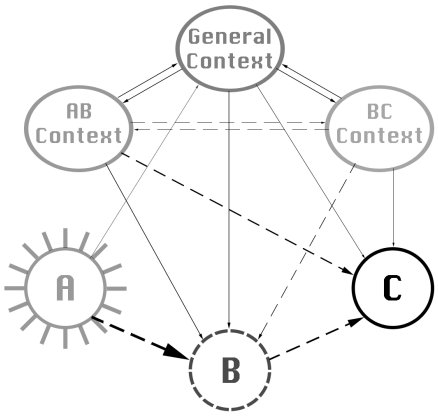

Figure 6. Network trained with the negative or inhibitory A-B and B-C contingencies that were presented in Treatment NN. The arrows represent connection weights that developed during training. Full arrows represent positive connections, and dotted arrows represent negative connections. The relative strength of each connection is indicated by the width of the arrow. In addition to the connection weights, the activation level of every unit during a test in which only unit A is turned on is shown. The radial lines around unit A indicate that only unit A was turned on (i.e., received external input). The nuance of each unit indicates how strongly it was activated during this test, with darker shades indicating stronger activation. Unit B was inhibited, as indicated by its dotted contour.

Some Conclusions

Yesterday, upon the stair,
I met a man who wasn't there,
He wasn't there again today,
I wish, I wish he'd go away...
(*Antigonish*, Hugh Mearns circa 1899; McCord, 1955)

One of us (AGB) is frequently reminded of the poem *Antigonish* when reading the various instantiations of propositional/ inferential theories. Inferences, beliefs and goals are things that we believe psychology should explain and are not things that should be used to explain behaviors. We acknowledge that this is an article of faith. The level of analysis for research is a meaningful epistemological question for all sciences and certainly is for psychology. The question is: What are the fundamental primitives of an appropriate explanation in psychology? Clearly, the early associationists believed the fundamental primitives were at the level of associations. Inferential and propositional theorists seem to believe that they are at the level of propositions or inferences, although, as we have mentioned, they are rarely clear about the exact form and constraints on these propositions. Our impressions are that, as a theory, they arise *deux ex machina* to explain findings that are difficult to account for with more reductionist explanations. It seems that these explanations involve a homunculus that has many of the properties of the cognitive processes that we wish to explain.

It would seem from this opening statement that we are unsympathetic to the various propositional theories, but we are not. These theories have generated a great deal of interesting empirical work and have discovered phenomena that might never have been investigated from a purely associationist perspective.

Moreover, many of these findings are at the moment very difficult to explain with moderately parsimonious associative networks. But it is the nature of science that many original mysteries defy reductionist explanations.

So how do we reconcile propositional accounts and associative accounts? We have argued that the levels of analysis outlined by Marr (1982) produce a useful perspective on this problem. He argued that a primary level of analysis is computation. Research asks just what the organism "computes" in the world. That is to say, what aspects of the physical world is the organism sensitive to and what is the nature of the response to these aspects? The world of physics includes rules about the behavior of objects and these can be written in mathematical or propositional from. No one would argue that, because a falling body follows the propositions of physics, it has instantiated them in its nature. Likewise, the world of cognition implies a series of rules and propositions. As we have mentioned before, these rules include descriptions of the rules of causal inference but also include the rules relating the events in conditioning (see also, Baker, Murphy, & Vallée-Tourangeau, 1996).

Thus, when a researcher shows that an organism's behavior maps onto a syllogism or other propositional structure, there are two possibilities of what this means. First, and we believe noncontroversially, it shows that the organism is sensitive to these rules about the world and presumably this sensitivity is an adaptation to accommodate them in behavior. But second, the organism may have an internal representation or algorithm that directly represents the rules of the world and this is what generates the behavior. Alternatively, some other mechanism, possibly associative, generates the computation. While it is obviously our position that the best and most parsimonious algorithm involves an associative approach, this does not mean we are correct. It is possible that the fundamental primitives necessary to explain causal reasoning and conditioning involve propositions. But, if so, it is crucial that proponents of this position generate theories with constraints that are potentially falsifiable. However, even if this is not done, the propositional approach has historically been, and still is, a useful heuristic for generating research.

The three research sections we have presented here illustrate this position. We initiated the research on choosing the most informative context from principles derived from causal model and power theory (Cheng, 1997; Lagnado, Waldmann, Hagmayer, & Sloman, 2007). The results were at least partially consistent with the theory but the finding that participants did not abandon the less informative contexts was not – unless new and unexpected propositions were generated to explain them. In the experiments on blocking of conditioned inhibition, we tested the argument that blocking would be more effective if the blocked stimulus could be unambiguously shown to have no effect or causal power. We found that this prediction of one version of inferential learning theory, in our hands at least, was not confirmed. And all of our results were broadly consistent with Rescorla and Wagner's (1972) associative model. But again we showed that an unconstrained inferential approach could accommodate the findings. In the final experiments on building causal chains from their links, we generated the predictions concerning the polarity and strength of the chains from formal logic and then verified them computationally with probability or contingency theory using the notions of causal

power and its derivative Δp (Cheng, 1997). We found that people's behavior mapped onto this analysis very well. We then showed how a simple associative net could account for this "propositional" behavior including the critical elements of timing and causal precedence.

We would be remiss in not pointing out that we have used the notion of parsimony rather cavalierly throughout. As Mitchell et al. (2009) have pointed out, associative theories are not necessarily parsimonious if for every new problem a new theory is generated. And, although our auto-associator has only six units in it, it does have many links. However, it should be emphasized that it computes predictions about timing and event sequencing. We are also trying to extend its use to a wider range of phenomenon but leaving it largely unchanged. Nonetheless, it should be emphasized that we are not immune to the parsimony argument we have used against the propositional accounts as algorithms. Nevertheless, it is our position that at least in terms of face validity the associative models are more plausible candidates for implementing in the physiology of the organism.

In conclusion we have argued that propositional accounts of cognition are very useful for generating research. They provide a useful framework for formalizing the rules of the world and asking what behavioral adaptations an animal might have that map onto them. However, for them to provide a useful theory or algorithm of behavior they must be more formally specified and be clearly falsifiable. Until this is done we still are concerned with the "… man who wasn't there."

References

Aitken, M. R., & Dickinson, A. (2005). Simulations of a modified SOP model applied to retrospective revaluation of human causal learning. *Learning & Behavior, 33*, 147-159.

Baetu, I. (2009). Associative and inferential accounts of extinction and blocking in causal learning. Ph.D. dissertation, McGill University, Montreal, Quebec, Canada. *Dissertation Abstracts International* (Publication No. AAT NR66377).

Baetu, I., & Baker, A. G. (2009). Human judgments of positive and negative causal chains. *Journal of Experimental Psychology: Animal Behavior Processes*, *35*, 153-168.

Baetu, I., & Baker, A. G. (2010). Extinction and blocking of conditioned inhibition in human causal learning. *Learning & Behavior*, *38*, 394-407.

Baker, A. G. (1977). Conditioned inhibition arising from a between-sessions negative correlation. *Journal of Experimental Psychology: Animal Behaviour Processes*, *3*, 144-155.

Baker, A. G., Murphy, R. A., & Vallée-Tourangeau, F. (1996). Associative and normative models of causal induction: Reacting to versus understanding cause. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation,* Vol. 34, (pp. 1-45). San Diego, CA: Academic Press.

Barberia, I., Baetu, I., Sansa, J., & Baker, A. G. (2010). Choosing optimal causal backgrounds for causal discovery. *Quarterly Journal of Experimental Psychology*, *63*, 2413-2431.

Beckers, T., De Houwer, J., Pineño, O., & Miller, R. R. (2005). Outcome additivity and outcome maximality influence cue competition in human causal learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *31*, 238-249.

Beckers, T., Miller, R. R., De Houwer, J., & Urushihara, K. (2006). Reasoning rats: Forward blocking in Pavlovian animal conditioning is sensitive to constraints of

causal inference. *Journal of Experimental Psychology: General, 135*, 92-102.
Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learning & Memory*, *11,* 485-494.
Brewer, W. F. (1974). There is no convincing evidence for operant or classical conditioning in adult humans. In W. B. Weiner & D. S. Palermo (Eds.), *Cognition and the symbolic processes* (pp. 1-42). Hillsdale, NJ.: Erlbaum.
Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367-405.
Chomsky, N. (1959). A review of B. F. Skinner's verbal behavior. *Language*, *35*, 26-58.
Cornsweet, T. N. (1970). *Visual perception*. New York: Academic Press.
De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & Behavior, 37*, 1-20.
De Houwer, J., Beckers, T., & Glautier, S. (2002). Outcome and cue properties modulate blocking. *Quarterly Journal of Experimental Psychology*, *55A*, 965-985.
De Houwer, J., Beckers, T., & Vandorpe, S. (2005). Evidence for the role of higher order reasoning processes in cue competition and other learning phenomena. *Learning & Behavior*, *33*, 239-249.
Dickinson, A., & Balleine, B.W. (1994). Motivational control of goal-directed action. *Animal Learning and Behavior, 22*, 1-18.
Fodor, J., & Piattelli-Palmarini, M. (2010). *What Darwin got wrong*. New York: Farrar, Straus, & Giroux.
Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior*, *4*, 267-272.
Hull, C. L. (1943). *Principles of behavior: An introduction to behavior theory*. New York: Appleton-Century-Crofts.
Hughes, D., Davis, J. D., & Grice, G. R. (1960). Goal box and alley similarity as a factor in latent extinction. *Journal of Comparative and Physiological Psychology*, *53*, 612-614.
Kamin, L. J. (1969). Selective association and conditioning. In W. K. Honig & N. J. Mackintosh (Eds.), *Fundamental issues in associative learning* (pp. 42-64). Halifax, Nova Scotia, CA: Dalhousie University Press.
Koehler, D. J., & James, G. (2009). Probability matching in choice under uncertainty: Intuition versus deliberation. *Cognition, 113,* 123-127.
Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation: Cues to causal structure. In A. Gopnik & L. Schultz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 154-172). New York: Oxford University Press.
Lovibond, P. F., & Shanks, D. R. (2002). The role of awareness in Pavlovian conditioning: Empirical evidence and theoretical implications. *Journal of Experimental Psychology: Animal Behavior Processes*, *28*, 3-26.
Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: Freeman.
Matzel, L. D., Held, F. P., & Miller, R. R. (1998). Information and expression of simultaneous and backward associations: Implications for contiguity theory. *Learning and Motivation, 19,* 317-344.
McClelland, J. L., & Rumelhart, D. E. (1988). *Explorations in parallel distributed processing: A handbook of models, programs, and exercises.* Cambridge, MA: MIT Press.
McCord, D. T. W. (1955). *What cheer: An anthology of American and British humorous and witty verse.* New York: The Modern Library.
Melchers, K. G., Wolff, S., & Lachnit, H. (2006). Extinction of conditioned inhibition

through nonreinforced presentation of the inhibitor. *Psychonomic Bulletin & Review*, *13*, 662-667.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, *32*, 183–246.

Murphy, R. A., & Baker, A. G. (2004). A role for CS-US contingency in Pavlovian conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, *30*, 229-239.

Murphy, R. A., Mondragon, E., & Murphy, V. A. (2008). Rule learning by rats. *Science*, *319*, 1849-1851.

Newton, I. (1934). Sir Isaac Newton's mathematical principles of natural philosophy and his system of the world. Berkeley, CA: University of California Press.

Pavlov, I. (1927). *Conditioned reflexes*. London: Oxford University Press.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current theory and research* (pp. 64-99). New York: Appleton-Century-Crofts.

Skinner, B. F. (1957). *Verbal learning*. New York: Appleton-Century-Crofts.

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgment. *Quarterly Journal of Experimental Psychology, 37B*, 1-21.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science, 27,* 453-489.

Tolman, E. C. (1932). Purposive behavior in animals and men. New York: Appleton-Century.

Tolman, E. C. (1948). Cognitive maps in rats and men. *The Psychological Review*, *55*, 189-208.

Tolman, E. C., & Honzik, C. H. (1930). "Insight" in rats. *University of California Publications in Psychology*, *4*, 215-232.

Vallée-Tourangeau, F., Murphy, R. A., Drew, S., & Baker, A. G. (1998). Judging the importance of constant and variable candidate causes: A test of the power PC theory. *Quarterly Journal of Experimental Psychology, 51A*, 65-84.

Vandorpe, S., De Houwer, J., & Beckers, T. (2005). Further evidence for the role of inferential reasoning in forward blocking. *Memory & Cognition*, *33*, 1047-1056.

Wagner, A. R. (1981). SOP: A model of automatic memory processing in animal behavior. In N. E. Spear & R. R. Miller (Eds.), *Information processing in animals: Memory mechanisms* (pp. 5-47). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Waldmann, M. R., Hagmayer, Y, & Blaisdell, A. P. (2006). Beyond the information given: Causal models in learning and reasoning. *Current Directions in Psychological Science*, *15*, 307-311.

Wu, M., & Cheng, P. W. (1999). Why causation need not follow from statistical association: Boundary conditions for the evaluation of generative and preventive causal powers. *Psychological Science*, *10*, 92-97.

Appendix

Like a few other associative models (e.g., Aitken & Dickinson, 2005; Wagner, 1981), this model represents events in real time. When a stimulus is physically present, the unit or units representing it receive some external input that causes their activation level to gradually increase. When the external stimulation ceases, the activation level of the units will gradually decay back to a resting level of zero. An active unit may spread its activation to other units via its connections. Depending on whether these associations are excitatory or inhibitory, the active unit will excite or inhibit the units connected to it.

The connections in the network do not represent temporal relationships (as in the temporal coding hypothesis; Matzel, Held, & Miller, 1998). Instead, the model learns temporal relationships merely as a result of units being active at various points in time. For instance, if a stimulus (A) is presented for a brief period of time, the activation level of the unit or units that represent it gradually increases and then decays back to zero when the stimulus ceases. If a second stimulus (B) is presented before the activation level of A has decayed, then there is an opportunity for an association from A to B to form. If B is presented at a later period of time when the activation level of A is very low, then the opportunity for an association to form is lost. Thus, the model explains the effect of delays between a potential cause and an effect simply by allowing a stimulus representation to decay from memory once the stimulus is no longer perceived.

Connections might form between any two units in the network. The model uses the delta learning rule (also used in the model developed by Rescorla and Wagner, 1972) to change the strength of the associations. According to this rule, the change in the connection from unit A to unit B ($\Delta W_{A\text{-}B}$) is computed as follows:

$\Delta W_{A\text{-}B}$ = (external activation of B – internal activation of B) x total activation of A (Equation 1)

Each unit has two sources of input that contribute to its activation level: an external source and an internal source. The external activation of B refers to the input that the unit receives from perception while stimulus B is presented. The internal activation of B refers to the input that the unit receives when other units activate it if they have become associated with B. Thus, when B is presented for the first time, its external input is positive, but its internal input is zero because its presentation was unexpected (i.e., no other unit predicted that B would occur). Since the difference between the external and internal input to unit B is positive, this allows A to develop an association with B, but only if the activation of A is not zero. This is because the change in the association from A to B depends not only on how surprising the occurrence of B is [represented by the term (external activation of B – internal activation of B)], but also on the activation level of A.

Of direct relevance to learning directed chains of events, the model also predicts that the association from A to B (A→B) becomes stronger than the association from B to A (B→A) if A precedes B in training. The B→A association is weaker because by the time B occurs, unit A no longer receives external input. Thus, after training, A might be able to activate the representation of B through the A→B association, but B will not be able to activate the representation of A since the B→A association is weak.