Associative accounts of causal cognition

Mike E. Le Pelley, Oren Griffiths & Tom Beesley

School of Psychology, University of New South Wales, Sydney, Australia

Corresponding author:

Dr Mike Le Pelley

School of Psychology

University of New South Wales

Sydney NSW 2052

Australia

Tel: +61 2 9385 1294

Email: m.lepelley@unsw.edu.au

Abstract

Humans are clearly sensitive to causal structures—we can describe and understand causal mechanisms and make predictions based on them. But this chapter asks: Is causal learning always causal? Or might seemingly causal behavior sometimes be based on associations that merely encode the information that two events "go together," not that one causes the other? This associative view supposes that people often (mis)interpret associations as supporting the existence of a causal relationship between events; they make the everyday mistake of confusing correlation with causation. To assess the validity of this view, one must move away from considering specific implementations of associative models and instead focus on the general principle embodied by the associative approach—that the rules governing learning are general-purpose, and so do not differentiate between situations involving cause–effect relationships and those involving signaling relationships that are non-causal.

KEYWORDS: learning, causal, association, associative, covariation, conditioning, contingency, contiguity, causal order, Rescorla-Wagner model.

The roots of the associative approach to causal cognition can be traced back to the British Empiricist philosophers of the 17th and 18th century. Most influential among these was David Hume, who viewed the causal relationship as central to all understanding, describing causation as ‘the cement of the universe’. Hume (1740/1978) noted in his ‘A Treatise on Human Understanding’ that neither the senses nor reason can establish that one object (a cause) is connected together with another object (an effect) in such a way that the presence of one necessarily entails the existence of the other. In other words, we cannot observe the nature of the causal connection between two events. In the case of a falling stone, we are unable to directly perceive some law of gravity that causes the stone to fall. But despite our inability to see or prove that there are necessary causal connections, we continue to think and act as if we had knowledge of such connections. Hume proposed that our beliefs in the connectivity between causes and effects arise simply as a result of the constant conjunction (or association) of those events – that is, by the perception of regularities in our experience. Hume asks us to consider the case of a putative Adam, brought to life in the midst of the world and in ‘the full vigour of understanding’. Adam would behave as if he had no knowledge of causal relationships in the world. He would be unable to predict that putting his hand in a flame would cause pain, or that clouds anticipate the likely onset of rain. Whereas we, endowed with the same senses and faculties of reasoning, are unable to resist making these and countless other such predictions.

The critical difference between Adam and ourselves is experience of associations between events in the world; in particular, between causes and effects. Thus we have observed many examples of one object striking another and causing this second to move. Despite, as Hume notes, the causal connection between these events being unobservable, this constant conjunction gives rise to an expectation that one event will be followed by a second. Hume proposed that it is this expectation that is manifest in the mind as a necessary connection, or belief. Hence the idea of necessary connection or causality, which arises as a result of regularities of experience, has its root in the mind, and is then projected onto the world in the form of predictive knowledge.

This idea that causal knowledge derives from experience of statistical regularities between events in the world clearly has strong ties to the notion of associative learning: learning that events are associated with one another, on the basis of experience of their co-occurrence. As a result of over a century's scientific study of associative learning in humans and nonhuman animals (going back to Thorndike, 1898), we now have a reasonable grasp of the fundamental principles underlying associative learning. Encouraged by this corpus of knowledge, Dickinson, Shanks and Evenden (1984) suggested that the established principles of associative learning could usefully be brought to bear on the issue of causal learning; in essence, they argued that causal cognition could usefully be brought under the umbrella of associative learning. This suggestion prompted a wave of empirical studies that have attempted to test the validity of this framework. We shall argue in this chapter that while the results of this research were, at best, mixed, the most important outcome was a greater understanding of the complexities and nuances of causal learning and behaviour.

Causal learning as associative learning

'Causal learning' and 'associative learning' are not synonyms. Crucially, associative learning simply requires a relationship to exist between the occurrence of two events, but *it does not require this relationship to be a causal one*. That is, associative learning is based on covariation, not causality. For example, Pavlov's (1927) dogs learned that a bell preceded delivery of food, but the sound of the bell did not *cause* the food to be delivered. Similarly, I might learn that the seminars in our department on Fridays tend to be more interesting than those on Wednesdays, but the day of the week does not cause one seminar to be better than another (that is, correlation does not equal causation).

In fact, we can divide examples of associative learning into three categories. The first involves instrumental (or operant) learning; that is, learning about the relationship between an action and the outcome that it produces. For example, flipping a light-switch is associated with a light turning on. Clearly, instrumental learning must involve a causal relationship.[1] The second category we can term causal Pavlovian learning. This involves the situation in which one external event signals the occurrence of a second external event by virtue of a causal relationship between them. For example, we might learn that when clouds appear in the sky, rain is likely to follow. The final category is non-causal Pavlovian learning, where one external event signals another but does not cause that second event to occur. The examples in the preceding paragraph of Pavlov's dogs and interesting seminar days fall into this latter category.

Recent work on the conditioning of attention provides a particularly clear demonstration of Pavlovian learning that is not based on causal knowledge. Le Pelley, Pearson, Griffiths and Beesley (2015) trained participants on a visual search task in which they were required to look at a target (a diamond) as quickly as possible. The appearance of a particular distractor stimulus (say, a red circle) in the search display on a given trial signalled that participants would receive a large monetary reward for looking at the target, while the presence of a different distractor (a blue circle) signalled that looking at the target would receive low reward. However, if participants looked at the distractor at any point prior to looking at the target, the reward that was scheduled for that trial was omitted. Under these circumstances, participants were nevertheless more likely to look at the 'high-value' distractor (the red circle in this case) than the 'low-value' distractor, even though doing so meant that they were more likely to miss out on large rewards than small rewards. That is, the high-value distractor was more likely to capture attention than was the low-value distractor. The implication is that attentional capture in this task was modulated by the Pavlovian relationship between distractor colour and reward: the high-value distractor had a Pavlovian relationship with large reward (i.e., it signalled the availability of large reward), and as a consequence of this relationship it became more likely to capture attention in future. Notably, this conditioned capture of attention by the high-value distractor occurred *even if participants were explicitly informed that looking at the distractor caused omission of the reward*, both in initial instructions and following every trial on which reward was omitted (Pearson, Donkin, Tran, Most & Le Pelley, 2015). In fact, the magnitude of the conditioned effect was unaltered compared to the case in which participants were not explicitly informed of this causal relationship regarding omissions. So even when participants were fully aware of the true causal relationship present in the task, their conditioned attentional response was not influenced by this knowledge; instead their conditioned responding continued to follow the (non-causal) Pavlovian relationship between colours and rewards. The suggestion, then, is that this example of a conditioned attentional response—wherein stimuli that predict high reward are more likely to capture attention than those that signal low reward—reflects the operation of a relatively automatic Pavlovian process that is not sensitive to causal knowledge.

To summarise, associative learning can be based on learning about causal relationships, but it can also be insensitive to causality. One view, then, is that causal learning is simply a subset of associative learning. According to this framework, there is a general-purpose set of principles that underlie associative learning, and these principles apply equally to situations involving causality and those that merely involve signalling (that is, covariation). The alternative is that there is something 'special' about causal relationships that sets them apart from instances of mere covariation (Cheng & Lu, this volume; Rottman, this volume; Waldmann & Hagmayer, 2005). To the extent that this alternative hypothesis is true, the associative account of causal learning will be found wanting. Below we consider some of the evidence for parallels between causal and non-causal associative learning.

Contiguity, contingency and cue competition

In a series of papers, Dickinson, Shanks and colleagues (for reviews, see Dickinson, 2001; Shanks, 1995) investigated whether the formation of causal beliefs in humans was subject to certain, 'standard' associative principles that had been established in earlier studies of animal conditioning. The first of these is the temporal contiguity of events; that is, the delay between their occurrence. Animal studies of non-causal Pavlovian conditioning show that the greater the delay between two events, the weaker the association that is formed between them. For example, if a tone CS reliably signals a puff of air to the eye (US), then a rabbit will learn to move its nictitating membrane (third eyelid) in response to the tone so as to attenuate the influence of the air puff. This conditioning occurs rapidly if the US follows 250 ms after the CS during training, but the rate of conditioning steadily decreases as the CS–US interval used during training becomes longer (Schneiderman & Gormezano, 1964; for related examples, see Gibbon, Baldock, Locurto, Gold & Terrace, 1977; Hawkins, Carew & Kandel, 1986; Ost & Lauer, 1965). Shanks and Dickinson (1991) demonstrated that causal learning in humans shows a similar sensitivity to temporal contiguity. In their task, participants could press a button whenever they liked, at a cost of 1 point. The outcome was the illumination of a figure on the screen, and every time this outcome occurred, participants gained 3 points. Participants were asked to maximise the number of total points gained. In fact, 90% of button-presses generated an outcome, so the optimal strategy was to press the button as frequently as possible. If the outcome immediately followed the action, then participants did indeed learn to press the button at a high rate, and at the end of the experiment reported a strong belief in a causal relationship between action and outcome. However, training with a delay of 2 or 4 s between action and outcome disrupted this pattern, producing a rate of button-pressing that did not differ from that of other participants trained under a regime in which pressing the button had no influence at all on whether the outcome occurred. This effect of delay was also reflected in participants' judgments of the strength of the causal relationship between action and outcome; with a 4-s delay, this relationship was judged to be around half as strong as with no delay.

Causal and non-causal learning also show similar sensitivity to the degree of contingency between events, where contingency is defined as the difference between the probability of event 2 (E2) occurring given that event 1 (E1) has occurred [denoted $P(\text{E2}|\text{E1})$], and the probability of E2 occurring in the absence of E1 [denoted $P(\text{E2}|\neg\text{E1})$]. Using rats, Rescorla (1968; see also Rescorla, 1967) demonstrated that the degree of contingency between a tone CS and an electric shock US modulated the strength of Pavlovian conditioning. One group of rats experienced training in which shocks only ever occurred in the presence of the tone CS. These rats showed evidence of rapidly developing a fear of the tone (in that hearing the tone would cause them to stop pressing a lever that delivered food). A second group of rats experienced the same number of occasions on which the CS and US were paired (with the same degree of temporal contiguity). However, for this group, additional USs could occur when the CS was not present, such that overall there was no contingency between CS and US, i.e., $P(\text{shock}\,|\,\text{tone}) = P(\text{shock}\,|\,\neg\,\text{tone})$. These rats showed no evidence of developing fear of the CS, suggesting that the lack of contingency between CS and US prevented the learning of an association between them. Shanks and Dickinson (1991; see also Chatlosh, Neunaber & Wasserman, 1985) demonstrated that causal learning of action–outcome relationships by humans is also sensitive to contingency. Using the procedure outlined in the previous paragraph, they found that participants' rate of button-pressing and judgments of the strength of the causal relationship between action and outcome declined systematically as the probability of unpaired outcomes (i.e. illuminations of the figure when the button had not been pressed) increased.

In fact, the relationship between contingency and learning is more subtle, and more interesting, than this. Studies of non-causal Pavlovian conditioning have shown that, when exposed to non-contingent presentations of a tone and shock, rats initially develop fear of the tone, but as experience continues this fear diminishes and eventually disappears (Rescorla, 1972). Moreover, the strength of preasymptotic fear depends on the overall frequency of shock: Rescorla also showed that this fear was greater in a condition in which $P(\text{shock}\,|\,\text{tone}) = P(\text{shock}\,|\,\neg\,\text{tone}) = 0.4$ than in a condition in which $P(\text{shock}\,|\,\text{tone}) = P(\text{shock}\,|\,\neg\,\text{tone}) = 0.1$. This is known as the *outcome density bias*. Both of these findings are mirrored in studies of causal learning in humans (for a review, see Shanks, 1995). Specifically, people's judgments of the strength of a relationship between a cause and an effect are influenced by the overall frequency of the outcome (the effect), even if there is no contingency between the 'cause' and the effect (Allan & Jenkins, 1983). However, as training continues this influence of outcome frequency on causal judgments decreases, with judgments eventually coming to reflect normative contingencies (Shanks, López, Darby & Dickinson, 1996; Wasserman, Chatlosh & Neunaber, 1983).

****
TABLE 1 ABOUT HERE
****

Further research shows that learning about the relationship between E1 and E2 is influenced not only by the degree of contingency between E1 and E2, but also by the degree of contingency between other events (E3, E4 etc) and E2, if these other events sometimes occur at the same time as E1. This is most neatly illustrated by the case of *blocking* (Kamin, 1968). A particularly clear, within-subjects demonstration of blocking of non-causal Pavlovian conditioning in animals is provided by McNally and Cole (2006); see Table 1. Rats were initially exposed to pairings of a visual CS, denoted stimulus A, with shock. In a subsequent phase of training, rats experienced trials on which A was presented in compound with a novel auditory stimulus, B, and again followed by shock. Also occurring in this latter phase were trials on which a compound of a novel visual stimulus, C, and a novel auditory stimulus, D, was paired with shock. Following this treatment, rats showed evidence of greater fear of D than of B, suggesting that they had formed a stronger D–shock association than their B–shock association. This is interesting, because the experienced contingency between B and shock is exactly the same as that between D and shock: both stimuli were paired with shock on the same number of trials, and shock was experienced in the absence of each stimulus the same number of times. Moreover, the temporal contiguity between CS and shock is the same for all CSs. This demonstrates that associative learning is not entirely determined by contingency and contiguity. In particular, it shows that associative learning about cue–outcome relationships does not proceed independently for each cue considered in isolation: B and D, considered in isolation, both have the same relationship with shock, and yet more is learned about D. Instead the implication is that learning in Stage 1 that A predicts shock, acts to block subsequent learning that B predicts shock on AB → shock trials in Stage 2. In contrast, since neither C nor D has been previously established as a predictor of shock when CD → shock trials are first experienced in Stage 2, neither will block learning about the other.

Blocking shows that cues interact in the learning process, and seem to compete for a limited amount of a learning resource that the outcome can support – hence blocking is often described as an example of *cue competition* in associative learning. And notably for the purposes of this chapter, blocking also occurs in human causal learning (e.g., Aitken, Larkin & Dickinson, 2000; Griffiths & Le Pelley, 2009; Griffiths & Mitchell, 2008; Le Pelley, Beesley & Griffiths, 2014; Le Pelley, Beesley & Suret, 2007; Shanks, 1985). For example, Aitken et al. used the common ‘allergy prediction’ paradigm, in which participants play an allergist whose aim is to discover the cause of reactions in a fictitious patient, Mr. X. On each trial, the participant is told the contents of a meal eaten by Mr. X, and must predict whether or not he suffered an allergic reaction as a result. Immediate corrective feedback is provided, and each different meal is encountered many times, which allows participants to learn the correct response for each meal. For example, they might learn that eating beef and sprouts reliably results in an allergic reaction, but eating bananas is not followed by a reaction. Aitken et al.’s Experiment 2 contained a blocking contingency and its control. In the blocking contingency, Stage 1 trials on which food A was always paired with allergic reaction (denoted A+ trials) were followed by Stage 2 trials on which a compound of foods A and B was always paired with allergic reaction (AB+). In the control contingency, participants experienced Stage 2 trials on which a compound of foods C and D caused reaction (CD+), but neither had previously been experienced in Stage 1. In a subsequent test phase participants were presented with each cue individually and asked to rate the efficacy of that food as a cause of allergic reaction. Blocking was demonstrated, in that food B was perceived as a weaker cause of allergy than was food D.

Another demonstration of an interaction between cues in learning comes from studies of the effect of signalling noncontingent outcomes. Rescorla (1972) showed that delivering shocks in the inter-trial intervals (ITIs) between pairings of a tone and shock resulted in reduced fear of the tone. This is not surprising, since these ITI-shocks result in a reduced contingency between tone and shock (see earlier). However, for a further group of rats, all ITI-shocks were signalled by a different cue (a clicker). Even though this signal does not alter the contingency between tone and shock, rats in this latter group showed significantly greater fear of the tone than did those for whom the ITI-shocks were unsignalled (see also Durlach, 1983; Rescorla, 1984). Similarly, signalling noncontingent occurrences of an outcome results in higher judgments of the strength of a causal relationship between an action (pressing a button) and an outcome (illumination of a figure on a screen) in humans (Shanks, 1986, 1989).

So, causal and non-causal learning show similar sensitivities to a range of factors. Both are influenced by the degree of contiguity and contingency between events, and both are subject to cue competition. These findings seem to support the idea that there is nothing ‘special’ about *causal* learning; that instead there is a set of general principles for associative learning, and that these principles apply equally to situations involving causality as those that don’t.

At first glance, then, the case for an associative account of causal learning seems to be on solid ground. However, this initial promise has subsequently been met with criticisms that can be broadly divided into two categories: those that question specific implementations of associative principles, and those that question associative principles in general. We shall argue that the former present no challenge to the associative account, whereas the latter force us to concede (unsurprisingly) that such an account must fall short of providing a full account of human causal reasoning.

The ‘If it’s not Rescorla–Wagner, it’s not associative’ Fallacy: General principles versus specific theories

Many of the phenomena of associative learning described above follow naturally from one of the most influential and famous theories of associative learning, the Rescorla–Wagner model (Rescorla & Wagner, 1972). At the heart of this model is the idea that the amount that is learned about a given cue on a given trial is related to the ‘surprisingness’ of the outcome on that trial. Formally, the model states that the strength of the association between a cue X and an outcome (denoted V_X) is updated on each trial according to the expression:

$$\Delta V_X = \alpha_X \, \beta_{US} \, (\lambda - \sum V) \qquad (1)$$

where ΔV_X represents the change in V_X on the current trial, and α_X and β_{US} are fixed parameters representing the salience of cue X and the outcome respectively. The error term ($\lambda - \Sigma V$) represents the discrepancy between the observed magnitude of the outcome (λ) and the magnitude of the outcome expected on the basis of all currently presented cues (ΣV); that is, how surprising the occurrence of the outcome is, given the presence of the presented cues.

Let us consider how this applies to a blocking contingency in which A+ trials are followed by AB+ trials. On initial A+ trials, the outcome is surprising, since it is not predicted by A; formally, we start with $V_A = 0$, and the magnitude of the outcome on this trial is λ (where $\lambda > 0$) so the error term on the first trial will be $\lambda - 0 = \lambda$. Hence an association will develop between A and the outcome; i.e., V_A will increase, and will continue to increase until (asymptotically) $V_{\mathrm{A}} = \lambda$. Consider now the AB+ trials of Stage 2. The outcome occurring on these trials is not surprising, since it is well-predicted by the presence of A; formally, on the first Stage 2 trial we have $V_A \approx \lambda$ (if we assume that Stage 1 training approaches asymptote) and $V_B = 0$ (since it is novel), and so the error term is given by $\lambda - \Sigma \mathrm{V} = \lambda - (V_A + V_B) \approx 0$. Hence little is learned on these trials, and consequently B does not form a strong association with the outcome. So the model correctly predicts that prior training with A will block later learning about B.

The Rescorla–Wagner model also provides a natural explanation of the finding that signalling noncontingent outcomes results in perception of a stronger cue–outcome relationship (Rescorla, 1972, 1984; Shanks, 1986, 1989). If we denote the critical cue as A, and the experimental context as C, then the condition in which noncontingent outcomes are *not* signalled can be thought of as involving AC+, C+ and C– trials (where the latter represents periods during which the subject is exposed to the context but outcomes do not occur). The context will develop a (weak) association with the outcome as a consequence of C+ trials. This will render the outcome occurring on AC+ trials less surprising than would otherwise be the case, since the error term on these trials is $\lambda - (V_A + V_C)$ with $V_C > 0$. Hence (to an extent) learning about the context will block learning about the cue, A. Now, if the noncontingent outcomes are signalled by a different stimulus, S, then training involves AC+, SC+ and C– trials. The presence of the salient stimulus S on SC+ trials will reduce conditioning of the context C on these trials (through a process of overshadowing: see Rescorla & Wagner, 1972). Consequently in this signalled group, the context will not compete with and block learning about the cue A to the same extent as in the unsignalled group. Hence the model anticipates stronger learning about A in the signalled group than in the unsignalled group, and this is the effect observed empirically both in causal and non-causal learning.

Successes of the Rescorla–Wagner model in explaining phenomena of causal learning such as blocking, the effect of signalling, and the outcome density bias (Shanks, 1995, provides a clear description of this latter success) fuelled the view that causal learning might be understood in terms of the associative principles that are implemented by this model, or one like it. This view had an unfortunate consequence, in that it has sometimes been taken to imply that any effect observed in causal learning that does *not* follow from the Rescorla–Wagner model therefore undermines the associative account of causal learning more generally.

An example of this is provided by the phenomenon of *backward blocking*. Recall that blocking involves A+ trials in Stage 1 followed by AB+ and CD+ trials in Stage 2; judgments of the causal strength of B following such training are lower than for D. A number of experiments have shown that a similar effect can be obtained even if the order of the training stages is reversed; i.e., if initial training is with AB+ and CD+, and this is followed by training with A+, then ratings of the causal strength of B are lower than those for D (e.g., Le Pelley & McLaren, 2001; Shanks, 1985; Wasserman & Berglan, 1998). The implication of this finding is that A+ trials in Stage 2 lead to a decline in the causal strength of B, even though B is not experienced during this phase – effects such as backward blocking provide examples of the *retrospective revaluation* of cues.

Notably, the Rescorla–Wagner model cannot account for backward blocking. Implicit in the model is the idea that only presented cues can undergo changes in associative strength: the salience of cue X (α_X) is assumed to be zero if the cue is not presented, and hence according to Equation (1), V_X cannot change if X is absent. So the model cannot explain the change in beliefs about B that must occur during A+ training in Stage 2 of a backward blocking procedure.

The fact that backward blocking lies beyond the Rescorla–Wagner model (and the alternative 'acquisition-based' models of associative learning that were dominant at the time, e.g., Mackintosh, 1975; Pearce & Hall, 1980; Wagner, 1981) led Shanks and Dickinson (1987) to wonder whether such findings undermined the associative approach to causal inference. But this concern is premature. While backward blocking clearly undermines *the Rescorla–Wagner model* (amongst others) as a complete account of causal inference, it does not necessarily undermine all possible implementations of an associative account. Indeed, it has been noted many times that the Rescorla–Wagner model does not provide a complete account of associative learning phenomena, regardless of the issue of causality (see Le Pelley, 2004; Miller, Barnet & Grahame, 1995). The model remains influential because, despite its simplicity, it is able to account for a variety of phenomena of conditioning—it has excellent heuristic value—but it does not explain everything.

To reiterate, the occurrence of backward blocking undermines certain implementations of an associative account of causal learning (including the Rescorla-Wagner model), but that does not mean that it necessarily falls outside the scope of associative models more generally. Indeed, several researchers have proposed associative models that are able to account for this effect (e.g., Dickinson & Burke, 1996; Le Pelley & McLaren, 2001; Tassoni, 1995; Van Hamme & Wasserman, 1994). We should be clear here that we do not mean to endorse a particular associative account of backward blocking. We would not necessarily even argue that backwards blocking is best interpreted as the outcome of an associative process. Our point is merely that this finding is, *at least in theory,* amenable to an analysis in terms of associative learning, and hence does not provide strong evidence on which to evaluate the utility of associative accounts considered as a class.

At the risk of becoming repetitive, we will restate our general point here since it is critical to our aim of evaluating associative accounts of causal learning: 'associative learning' and 'the Rescorla–Wagner model' are *not* synonymous. Associative learning is a generic term referring to learning about the covariation between events based on experience with those events. The Rescorla–Wagner model is just one possible mechanism for implementing associative learning. In this chapter, when we contrast associative learning and causal learning, this does not commit us to a particular mechanistic view of how associative learning occurs. For example, one view (typified by the Rescorla–Wagner model) sees associative learning as mediated by the formation of physical links between mental representations of stimulus elements. In some models, formation of these links is mediated by the attention that is paid to events (e.g., Le Pelley, 2004, 2010; Mackintosh, 1975; Pearce & Hall, 1980). An alternative, exemplar-based approach instead assumes that it is configurations of elements that form links with other events (e.g., Kruschke, 1992, 2003; Medin & Schaffer, 1978). Yet another view eschews the idea of links altogether, and instead argues that associative learning is mediated by calculation of probabilities (e.g., Cheng & Novick, 1990; Peterson & Beach, 1967), or by propositions that describe the associative relationship between events (Mitchell, De Houwer & Lovibond, 2009). A large body of fascinating research has been conducted in an attempt to decide between these alternative mechanisms for associative learning (see, for example, the chapter by Boddez, De Houwer & Beckers, this volume), but this issue is orthogonal to the scope of this chapter. Instead, for the purposes of this chapter, 'associative learning' simply means 'learning about covariation, however this is implemented', and is contrasted with 'causal learning', which implies sensitivity to causal structure beyond merely covariation.[2]

Is there anything special about causality? Part 1: Causal order

At this point, the reader may be growing concerned about the issue of falsifiability. Might it be possible to come up with an associative model that could explain *any* pattern of data? If so, then we might never be able to find evidence that would allow us to judge the value of the associative approach considered in its most general terms, which would render it uninteresting from a research perspective. However, we shall argue that this is not the case. Specifically, we believe there are certain properties that should be common to any associative account, and which—when attempting to apply that account to studies of causal learning—generate testable and falsifiable predictions. The crucial point is that, as noted earlier, the associative account provides a general-purpose set of principles that apply equally to all learning situations, whether those situations involve causality or not. In other words, associative accounts are essentially blind to the nature of the events that are involved in learning. On this view, there is nothing ‘special’ about causal relationships that sets them apart from instances of mere covariation.

An example should make this clear. Associative accounts involve learning about signalling relationships based on temporal order; that the occurrence of *cue events* A and B signal that an *outcome event* X will occur subsequently. Hence according to any associative account, the ‘direction’ of learning—from cues to outcomes—is determined by the temporal order of events. In contrast, the direction of causality flows from causes to effects, regardless of the temporal order in which those events are encountered.

Imagine that participants experience a number of trials on which cues A and B appear together, and signal that outcome X will occur. Under these circumstances, we would expect cue competition between A and B for the limited amount of learning that the outcome X can support (similar to the case of blocking described earlier). That is, learning about the relationship between A and X will be weaker if training is with AB→X (in which competition between A and B will occur; the presence of B is said to *overshadow* learning about A, and vice versa), than if it is with A→X (in which case there will be no competition).

We can now contrast two cases. In the first, the scenario is such that cues A and B are naturally interpreted as causes (e.g., foods eaten by a patient) and X as an effect (an allergic reaction caused by eating allergenic foods). In this case the temporal order of events is aligned with their causal order; this is often referred to as a *predictive learning* scenario. In the second case, the scenario is such that cues A and B are naturally interpreted as effects (e.g., symptoms suffered by a patient) and X as a cause (a disease that produces those symptoms). In this latter case, the temporal order of effects (A and B precede X) is opposite to their causal order (X causes A and B); this is a *diagnostic learning* scenario, in that participants learn to diagnose the cause on the basis of the presence of symptoms.

The important point is that, to an associative model, these two cases are equivalent. In both, the model learns to predict X on the basis of the presence of A and B, regardless of the causal status of these events. Hence any phenomenon of learning, such as cue competition, that occurs in a predictive learning scenario should also be observed in an otherwise-comparable diagnostic learning scenario. However, if people are sensitive to the causal status of events, then we might expect to see differences between predictive and diagnostic scenarios, reflecting the different causal order inherent in each. In particular, Waldmann and Holyoak (1992, 1997) noted that one should expect competition between multiple independent causes of a common effect, but not between multiple independent effects of a common cause. If we have experience that eating apple and banana causes allergic reaction, and encounter independent evidence that apples cause allergy, then this should lead us to reduce the strength of our belief that banana causes allergy. However, if we have experience that suffering from Jominy Fever causes nausea and pustules, and encounter independent evidence that Jominy Fever causes nausea, this should not lead us to change our belief that it also causes pustules. In other words, if participants are sensitive to causal order, then cue competition should be observed in a predictive scenario but not in a diagnostic scenario.

The empirical evidence on this issue is mixed. Some studies have found evidence of an asymmetry, with cue competition observed in predictive but not diagnostic scenarios (e.g., Booth & Buehner, 2007; Tangen & Allan, 2004, Experiments 1 and 2; Van Hamme, Kao & Wasserman, 1993; Waldmann, 2000, 2001; Waldmann & Holyoak, 1992). These findings imply a sensitivity to causality; that there is something special about a causal relationship that distinguishes it from mere signalling of outcomes by cues. Such demonstrations of asymmetry therefore undermine associative accounts of causal learning. And, unlike backward blocking, this undermining applies not just to a specific implementation of an associative account, but to the approach in general. Having said that, however, a number of other studies have reported similar cue competition effects in both predictive and diagnostic scenarios (e.g., Cobos, López, Caño, Almaraz & Shanks, 2002; Le Pelley & McLaren, 2001; Matute, Arcediano & Miller, 1996; Shanks & López, 1996; Tangen & Allan, 2004, Experiments 3 and 4). The symmetry between conditions found in these latter studies implies a lack of sensitivity to causal order, and thus supports the general-purpose, cue→outcome approach taken by associative models of learning.

How, then, are we to reconcile these discrepant findings? A likely possibility relates to the ‘availability’ of the causal scenario to participants. In the studies cited above that found symmetry between predictive and diagnostic learning (suggesting insensitivity to causal order) participants simply read instructions regarding the particular causal scenario under which they were being tested. The importance of considering causal order was not made explicitly salient to them, and, perhaps as a result, the data suggest that they did not consider this information when making judgments. In contrast, in the studies demonstrating asymmetry, causal order was typically made more salient. For example, Waldmann (2000, 2001) had participants summarize the instructions prior to the experiment in order to verify that they had correctly understood the different causal structures involved in the different cover stories. It seems likely that this requirement emphasised to participants the importance of making use of causal order when forming their judgments, and as a consequence no cue competition was observed in the diagnostic condition.

More direct evidence that the salience of the causal structure is the critical determinant of (a)symmetry between predictive and diagnostic learning comes from a study by López, Cobos and Caño (2005). In Experiment 1A, participants read ‘standard’ instructions about the causal scenario, which involved an electrical box with lights on the front and back. In the predictive condition, participants were told that illumination of lights on the front caused illumination of the lights on the back; in the diagnostic condition, they were told that illumination of lights on the front was caused by the lights on the back. Participants then completed a cue competition task in which, on each trial, they were shown which lights were illuminated on the front of the box and were required to predict which bulbs would be lit on the back (with corrective feedback). Results showed near-perfect symmetry, with equal evidence of cue competition in predictive and diagnostic conditions. Importantly, this symmetry did not reflect participants’ failure to understand the task’s causal structure, since symmetry was also observed if analysis was restricted to only those participants who could remember and understand the causal structure in a comprehension test conducted after the cue competition task. Experiment 2 was exactly the same as Experiment 1A, except that a sentence was added to the initial instructions to emphasise the importance of making use of the information regarding causal structure: “To solve the task correctly, it is VERY IMPORTANT to take into account what you have just read in the instructions. Most importantly, in order to solve the different examples of the task… bear in mind that the lights on one side of the box cause the illumination of the lights on the other side of the box.” This instruction was sufficient to induce asymmetry, with significantly stronger cue competition in the predictive scenario than the diagnostic scenario. However, it did not improve participants’ understanding of the causal structure of the task as assessed in the final comprehension test, which was similar regardless of the inclusion or exclusion of the critical sentence in the instructions.

To summarize, in Experiment 1A López et al. (2005) showed that participants will not necessarily show sensitivity to causal structure even when they are aware of, and understand, this structure. In Experiment 2 López et al. showed that, given explicit prompting to use information about causal structure when making judgments, participants were able to do so. Here we follow Shanks (2007) in arguing that these data support a ‘dual-model’ approach to learning. The implication is that, given sufficient motivation, people can reason about causality in a non-associative way that distinguishes between predictive and diagnostic causal structures. This should come as no surprise whatsoever – the fact that we can write about, and you can understand, the distinction between these causal structures makes it clear that people are able to comprehend and make use of the difference between them. However, it also seems that, without prompting, people’s default approach is simply to learn about relationships between cues and outcomes in a manner that is insensitive to causal order. This default pattern operates in exactly the way anticipated by an associative account.

A caveat of sorts is required here, relating to differences in the ease of mapping between temporal orders and causal orders. Consider López et al.’s (2005) study, in which participants were required to use information regarding which lights were illuminated on the front of the box (cues) to predict which lights would be illuminated on the rear of the box (outcomes). In the predictive learning condition, participants were told that illumination of lights on the front (causes) caused illumination of lights on the rear (effects). Hence for these participants the temporal order of events (in which cues precede outcomes) was aligned with their causal order (in which causes precede effects), making it easy to map from temporal order to causal order by mapping cues to causes, and outcomes to effects. In contrast, in the diagnostic learning condition participants were told that illumination of lights on the front (effects) was caused by illumination of lights on the rear (causes). Hence in this condition the temporal order of events opposed their causal order. So diagnostic causal learning under these circumstances would require separating the temporal order of learning events from their causal order: it requires mapping cues to effects, and outcomes to causes. This separation of learning events and mental representations may be cognitively demanding. Perhaps participants might simply be unwilling to engage in this effortful process, and fall back on the simpler, predictive mapping (cues = causes, outcomes = effects) even though it is at odds with the actual causal structure. On this account, symmetry between diagnostic and predictive learning conditions—as observed in López et al.'s Experiment 1—does not necessarily reflect non-causal reasoning in the diagnostic condition. People may instead by reasoning causally, but incorrectly, in this condition.

At this point the problem of unfalsifiability looms once more. If we find that a participant's learning is at odds with causal structure, we could presumably always save a single-process, causal-only account by arguing that the participant is reasoning causally, but is doing so based on an incorrect causal structure (even if, as in López et al.'s study, the participant is fully able to report the correct causal structure when asked). We shall consider a case below which stretches the plausibility of this 'incorrect causal reasoning' account even further, but more generally we leave it up to the reader to decide from themselves whether this type of essentially unfalsifiable argument is satisfactory.

Is there anything special about causality? Part 2: Revisiting temporal contiguity

We noted earlier that a critical factor in the learning of cue→outcome associations (whether they are causal or non-causal) is the degree of temporal contiguity between cue and outcome. Typically, a shorter delay between cue and outcome produces stronger learning, and (in the case of a causal relationship) stronger judgments of causality (see Buehner, this volume). The word ‘typically’ is important here, because in fact there are cases in which this law of temporal contiguity is broken. As in the case of causal order, these exceptions arise as a result of people taking account of the causal context of the judgment they are being asked to make. And once again, this influence of the specific nature of the events involved in learning on the pattern of what is learned, runs counter to the general-purpose approach offered by associative accounts.

Consider, for example, a study by Buehner and May (2004; see also Buehner & May, 2002, 2003), in which participants were asked to judge whether pressing a light switch made a bulb illuminate. For all participants, there was a 75% chance of the bulb illuminating if the switch had been pressed. In the *zero delay* condition, this illumination would occur immediately; in the *long delay* condition the bulb would illuminate 4 s after the button was pressed. The bulb never illuminated if the switch was not pressed.

One group of participants was told that the bulb was an ordinary light bulb that should light up right away. These participants showed a standard temporal contiguity effect: after a training period during which they could press (or not press) the switch and observe the consequences, judgments of the strength of the causal relationship between switch and light were lower in the long delay condition than in the zero delay condition. In contrast, a second group of participants was told that the bulb was an energy-saving model that took 4 s to light up. These participants did not show a detrimental influence of delay on causal beliefs – ratings of causality were equally strong in the zero delay and long delay conditions. These results clearly demonstrate that a change in the causal model that underlies the task can change the influence of temporal contiguity on the beliefs that are developed. If the task was framed with a causal structure in which delay was expected, then causal judgments no longer suffered as a result of experience of a delay. The implication is that participants' understanding of the nature of events that they are experiencing influences the beliefs that they develop as a result of that experience. This finding is very difficult to reconcile with a purely associative account in which learning is about the relationship between cues and outcomes and is essentially blind to beliefs about the causal nature of the events that constitute those cues and outcomes.

It is worth reflecting on the finding that participants who expected a delay in Buehner and May's (2004) experiment were not sensitive to temporal contiguity (see also Buehner & May, 2002, 2003): that is, their judgments were equally strong in the zero delay and long delay conditions. Surely, if participants were making proper use of the causal model inherent in the task description, judgments should actually have been *lower* in the zero delay condition. If they expected the bulb to illuminate 4 s after pressing the switch, then an immediate illumination should have been perceived as an 'uncaused effect', and weakened the perception of causality. The fact that this did not happen might be taken to suggest that close temporal contiguity is sufficient to support development of a causal belief even in the face of contrary expectation, and this suggestion is clearly easier to reconcile with an associative approach. That said, if the causal structure of the task is made more available, in order to strengthen participants' expectation of a delay, then an advantage for training under long delay conditions relative to short delay is observed (Buehner & McGregor, 2006). This is somewhat similar to the manipulations used to strengthen participants' attention to causal order described in the previous section. As in that case, the unsurprising conclusion is that, given sufficient motivation to consider the causal structure of a task when making judgments, participants are able to do so.

A study by Schlottmann (1999) is particularly interesting in this regard. Children aged between 5 and 10 were presented with a 'mystery box'. A ball dropped into one of two holes in one end of the box would cause a bell to ring at the other end, either immediately or after a short delay (3 seconds), depending on the hidden mechanism inserted into the box. Initially, children received extensive experience with the two mechanisms outside the box, and were guided by the experimenter through a series of exercises designed to help them understand how one mechanism (a seesaw) would cause the bell to ring quickly when a ball was dropped, and the other (a runway) caused it to ring after a pause. This culminated in a prediction test in which, on each trial, one of the mechanisms was inserted into the box out of sight of the children. A ball was then dropped in, and children observed whether the bell rang immediately or after a delay; they were then asked to predict which of the mechanisms (seesaw or runway) was inside the box. Performance on this prediction test was near-perfect, regardless of the children's age, demonstrating that children understood the difference in timing produced by the two mechanisms, and could make inferences on the basis of this understanding.

In the final, critical test, children saw the experimenter place one of the two mechanisms in the box, but could not see which of the two holes the mechanism was placed under. A picture of the mechanism was placed in view on the outside of the box as a reminder. Next, the experimenter dropped one ball into one of the holes, paused, and then dropped a second ball into the second hole; the bell rang immediately after the second ball was dropped. Children were asked which of the two balls had made the bell ring. The correct answer to this question depends on the mechanism hidden inside the box. If the fast mechanism was present, then the correct answer would be that the second (contiguous) ball had made the bell ring. If the slow mechanism was present, then the correct answer would be the first (delayed) ball. Ten-year-olds were clearly able to make this distinction. However, younger children (5-7 year-olds) could not. In particular, when the slow mechanism was in the box, a majority of these younger children still claimed that the second (contiguous) ball had caused the bell to ring, even though these same children had demonstrated understanding that the slow mechanism produced a delay in the earlier prediction test.

These results suggest that while older children are able to incorporate their causal understanding into their judgments, younger children instead continued to be led by contiguity even when this was at odds with the causal structure of the task. One interpretation of this difference is that young children’s judgments are more likely to reflect the product of relatively automatic, associative processes that are insensitive to causal knowledge, whereas older children have developed the cognitive flexibility and executive function necessary to override these associative influences by the appropriate deployment of causal knowledge (given sufficient motivation to do so, and a sufficiently clear causal mechanism). This suggestion is certainly not without precedent in the developmental literature (e.g., see Kendler & Kendler, 1959, 1962, 1970; Kuhn & Pease, 2006).

At the end of the previous section, we considered the possibility that cases of insensitivity to true causal structure, rather than supporting a dual-process account in which learning is sometimes the product of non-causal, associative processes, may instead reflect causal reasoning based on an incorrect causal structure. While such an argument could *in theory* be extended to explaining the insensitivity to the true causal mechanism demonstrated by Schlottmann’s (1999) younger children, this stretches the bounds of plausibility. These children were able to understand the true causal structure, as demonstrated by their performance in the prediction test. And application of a causal structure involving a three-second delay between cause and effect does not seem like it should be significantly more cognitively demanding than a structure with no delay. Hence we would argue that insensitivity to causal structure under these circumstances provides good evidence that learning need not always be mediated by causal reasoning.

Selective conditioning in nonhuman animals

The discussion in the previous sections is based around the premise that associative models involve learning about relationships between cues and outcomes, but are blind to the nature of the events that constitute those cues and outcomes. In fact the argument needs to be more nuanced than this. This is because there are well-established demonstrations in the literature on animal conditioning showing that learning can be influenced by the nature of the stimuli involved.

Perhaps the clearest demonstration of this comes from classic studies of selective conditioning (also known as preparedness) in rats (e.g., Domjan & Wilson, 1972; Garcia & Koelling, 1966). For example, Domjan and Wilson trained two groups of rats. For one group, an infusion of saccharinflavoured water into the mouth was used as the CS; for the other group, the sounding of a buzzer was the CS. For both groups, the US that was delivered at the termination of the CS was injection with lithium chloride in order to induce nausea. After three conditioning trials, rats were given two preference tests. Each test featured two drinking tubes. In the saccharin preference test, one tube was filled with saccharin and the other with water. In the buzzer preference test, both tubes were filled with water, but a buzzer was activated whenever the rat licked at one of the two tubes. Results indicated that rats had learned to avoid the saccharin flavour, but not the buzzer: on saccharin preference tests rats drank more from the water tube than the saccharin tube, but on buzzer preference tests rats drank from the buzzer tube and non-buzzer tube equally. The implication is that rats had learned to associate the flavour, but not the sound, with illness. Crucially, this pattern was not simply a result of the flavour being a more intense or noticeable CS. Two further groups of rats were conditioned using the same CSs, but the US was now an electric shock. In the final preference test these rats showed avoidance of the buzzer tube, but not of the saccharin tube—the exact opposite of rats trained with the nausea US—suggesting that they had learned to associate the sound, but not the flavour, with shock.

This double dissociation clearly shows that the nature of the stimuli is an important determinant of conditioning in animals. Rats learned to associate an interoceptive CS (flavour) with an interoceptive US (nausea), and an exteroceptive stimulus (sound) with an exteroceptive US (shock), but did not learn to associate an interoceptive CS with an exteroceptive US or vice versa. So what, if any, is the fundamental difference between this example and the demonstrations of sensitivity to causal structure in humans described in previous sections? One option would be to argue that there is no difference: that examples of selective conditioning demonstrate that even the simplest examples of conditioning are a consequence of animals reasoning based on an internal causal model of the world (in which internal cues tend to produce internal outcomes, and external cues tend to produce external outcomes). If this is the correct interpretation, then the case for associative accounts of learning seems bleak.

However this interpretation is not one to which we subscribe. For one thing, the suggestion that even simple examples of Pavlovian conditioning reflect reasoning about causal structure has difficulty explaining the persistence of conditioned responding in the face of an omission contingency (see earlier). Instead it seems more likely that selective conditioning reflects innate predispositions that have been shaped by evolutionary processes (Rozin & Kalat, 1971). For example, in the evolutionary environment of a rat, tastes are likely to provide more reliable information than sounds regarding whether a food is poisonous. As such, animals who are predisposed to learn about such relationships will be more likely to survive and pass on this characteristic.

Strong evidence in favour of this evolutionary account of selective conditioning comes from studies suggesting that the effect is present from birth, and certainly before rats have had useful prior experience with the sorts of stimuli that are involved. Gemberling and Domjan (1982) conditioned rats when they were only 24 hours old with either lithium-induced nausea or shock. These newborn rats showed clear evidence of selective conditioning. When the US was nausea, the saccharin flavour constituted an effective CS but an exteroceptive texture cue (either a rough cloth or the smooth interior of a cardboard box) did not. In contrast, when the US was shock, rats showed stronger evidence of conditioning to the texture than to the flavour.

In summary, while selective conditioning in rats demonstrates a sensitivity to the nature of the events that are involved in conditioning, it seems unlikely that this is a consequence of the animals engaging in 'online' inferences based on an internalized causal model of the world. Instead this sensitivity seems to reflect innate predispositions towards associating certain classes of stimuli, but not others. Hence a general-purpose associative account is sufficient to explain these findings if it is augmented with the idea that this associative process operates through a filter of innate mechanisms that will tend to favour formation of some associations over others. This is quite different from the cases of sensitivity to causal structure in humans described earlier. Many of those experiments used scenarios in which it is implausible to suggest that we have developed innate evolutionary tendencies (e.g., balls rolling along runways and ringing bells; switches turning on light-bulbs; grenade-launchers firing at tanks etc). Instead it seems clear that sensitivity to causal structure in such studies results from a process of causal reasoning based on an internalized model of the world, with this reasoning taking place during the task itself. And it seems equally clear that such a process lies beyond any purely associative approach to learning.

Conclusions

In this chapter, we have argued that in attempting to assess the validity of an associative account of causal learning, we need to move away from considering specific implementations of associative models and instead focus on the general principle embodied by the associative approach. This principle is that learning is concerned with the relationship between cues and outcomes; that the rules governing learning are general-purpose, and hence do not differentiate between situations involving cause–effect relationships and those involving signalling relationships that are non-causal. An association merely encodes the information that two events 'go together', not that one *causes* the other. Ultimately, the associationist framework argues that causal learning can be based on a process that is fundamentally non-causal. In other words, people may encode an association between event E1 and event E2 that has no notion of causality about it. However, they may then interpret this association as supporting the existence of a causal relationship between events. Suppose that I have formed a strong association between E1 and E2, but a weak association between E3 and E2. I am then asked which of E1 or E3 is more likely to be a cause of E2. Under these circumstances it would seem a sensible heuristic to choose the event that has the stronger association with E2 (E1 in this case), even though the associations do not actually encode causal information.

This idea that associations have heuristic value with regard to causality is an important one. Events that are causally related will typically tend to go together in the world, and hence will tend to become associated. As such the existence of an association can reasonably be taken as an indication of a causal relationship. Of course, this inductive leap from observing correlation to inferring causation is actually logically invalid: as psychology undergraduates are repeatedly told, ‘Correlation does not imply causation’. But the fact that we need to repeat this mantra so often to our students is also an indication of how seductive it is to equate correlation and causation. Indeed, humans seem to be very susceptible to such correlational learning: superstitions (e.g., walking under a ladder will cause bad luck; carrying a rabbit’s foot will bring good luck) occur in all societies around the world, and can be seen as a failure to distinguish between causes and mere associates.

The implication is that, often, people are happy to take associations between events as a proxy for causal relations. Why should this be the case? The most likely reason, as is generally the case for heuristics, is one of effort. It is much easier to encode the covariation between two events than to impute a causal relationship between them, because the latter requires making interventions and observing their consequences (cf. Rottman, this volume; Waldmann & Hagmayer, 2005), but this may not always be easy or even possible. For example, it is simple to observe that the phase of the moon covaries with the size of ocean tides, but proving a causal relationship by intervening to change the moon’s phase and observing the effect on the tide is impossible. The effort argument becomes even more apparent as additional potential causes are introduced. It is simple to apply the Rescorla–Wagner model to any number of cues, presented individually or in combination, in order to track their covariation with any number of outcomes, again presented individually or in combination. However, deriving an accurate causal model of the relationships between multiple stimuli and outcomes on the basis of experience is a considerably more challenging computational problem (Chickering, Heckerman & Meek, 2004; Ellis & Wong, 2008).

Of course, this is not to suggest for a moment that people are *unable* to reason in a more careful fashion about causal relationships. We have provided examples in this chapter in which people have been shown to be sensitive to causal structures when making judgments in a way that does not follow from an associative approach. But really, these examples are unnecessary – it is immediately obvious that people can describe and understand causal mechanisms, and so it would be lunacy to suggest that an associative account could ever provide a full account of human causal behaviour. However, as Shanks (2007) noted, “To focus (as many researchers do) on the fact that there is a pattern of judgements… that is inconsistent with associative theory is to miss the point: To repeat, it is not a matter of debate that people can reason normatively or logically under certain circumstances” (p300). Of more interest (at least for the purposes of the current chapter) is the finding that judgments often do *not* reflect sensitivity to causal structure, even when participants can evidently understand the causal structure involved in the task. Such a pattern has typically been reported in situations in which participants are not given strong prompting to make use of causal structure information when making their judgments. These findings point to a role for associative processes in explaining some aspects of causal behaviour, and suggest more generally that, in some cases, causal judgments may be made based on the output of a process that is blind to the notion of causality.

Footnotes

[1] A possible exception is superstitious beliefs. For example, I might (by coincidence) experience several occasions on which I wear a particular pair of socks and my favourite football team wins their match, and as a result I might develop the instrumental belief that wearing those socks causes the team to win. But of course there is no real causal relationship between these events. Nevertheless, this form of superstitious instrumental learning is based on the *belief* in a causal relationship.

[2] There is a subtlety here that is worth noting. As mentioned, one view that has been proposed is that associative learning is mediated by the formation of propositions that describe the relationship between events (Mitchell et al., 2009). Unlike the other approaches to associative learning outlined here, which are restricted to representing covariation, the propositional account has the potential to represent *either* covariation information (i.e., the proposition that 'event A is associated with event B'), *or* causal information (the proposition that 'event A causes event B'). So learning based on propositions could be either causal or non-causal (associative, in the current terminology) in different situations.

References

Aitken, M. R. F., Larkin, M. J. W., & Dickinson, A. (2000). Super-learning of causal judgements. *Quarterly Journal of Experimental Psychology, 53B*, 59-81.

Allan, L. G., & Jenkins, H. M. (1983). The effect of representations of binary variables on judgment of influence. *Learning and Motivation, 14*, 381-405.

Boddez, Y., De Houwer, J., & Beckers, T. (this volume). The inferential reasoning theory of causal learning: Towards a multi-process propositional account. In M. R. Waldmann (Ed.), *The Oxford Handbook of Causal Reasoning*. Oxford, UK: Oxford University Press.

Booth, S. L., & Buehner, M. J. (2007). Asymmetries in cue competition in forward and backward blocking designs: Further evidence for causal model theory. *Quarterly Journal of Experimental Psychology, 60*, 387-399.

Buehner, M. J. (this volume). Space, time, and causality. In M. R. Waldmann (Ed.), *The Oxford Handbook of Causal Reasoning*. Oxford, UK: Oxford University Press.

Buehner, M. J., & May, J. (2002). Knowledge mediates the timeframe of covariation assessment in human causal induction. *Thinking & Reasoning, 8*, 269-295.

Buehner, M. J., & May, J. (2003). Rethinking temporal contiguity and the judgement of causality: Effects of prior knowledge, experience, and reinforcement procedure. *Quarterly Journal of Experimental Psychology, 56A*, 865-890.

Buehner, M. J., & May, J. (2004). Abolishing the effect of reinforcement delay on human causal learning. *Quarterly Journal of Experimental Psychology, 57B*, 179-191.

Buehner, M. J., & McGregor, S. (2006). Temporal delays can facilitate causal attribution: Towards a general timeframe bias in causal induction. *Thinking & Reasoning, 12*, 353-378.

Chatlosh, D. L., Neunaber, D. J., & Wasserman, E. A. (1985). Response-outcome contingency: Behavioral and judgmental effects of appetitive and aversive outcomes with college students. *Learning and Motivation, 16*, 1-34.

Cheng, P. W., & Lu, H. (this volume). Consideration of constraints necessary for creating a causal representation of the world. In M. R. Waldmann (Ed.), *The Oxford Handook of Causal Reasoning*. Oxford, UK: Oxford University Press.

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology, 58*, 545-567.

Chickering, D. M., Heckerman, D., & Meek, C. (2004). Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research, 5,* 1287-1330.

Cobos, P. L., López, F. J., Caño, A., Almaraz, J., & Shanks, D. R. (2002). Mechanisms of predictive and diagnostic causal induction. *Journal of Experimental Psychology: Animal Behavior Processes, 28,* 331-346.

Dickinson, A. (2001). Causal learning: An associative analysis. *Quarterly Journal of Experimental Psychology, 54B,* 3-25.

Dickinson, A., & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgements. *Quarterly Journal of Experimental Psychology, 49B,* 60-80.

Dickinson, A., Shanks, D., & Evenden, J. (1984). Judgment of act-outcome contingency: The role of selective attribution. *Quarterly Journal of Experimental Psychology, 36A*, 29-50.

Domjan, M., & Wilson, N. E. (1972). Specificity of cue to consequence in aversion learning in the rat. *Psychonomic Science, 26,* 143-&.

Durlach, P. J. (1983). Effect of signaling intertrial unconditional stimuli in autoshaping. *Journal of Experimental Psychology: Animal Behavior Processes, 9*, 374-389.

Ellis, B., & Wong, W. H. (2008). Learning causal Bayesian network structures from experimental data. *Journal of the American Statistical Association, 103,* 778-789.

Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science, 5*, 121-122.

Gemberling, G. A., & Domjan, M. (1982). Selective associations in one-day old Rats: Taste-toxicosis and texture-shock aversion learning. *Journal of Comparative and Physiological*

Psychology, 96, 105-113.

Gibbon, J., Baldock, M. D., Locurto, C., Gold, L., & Terrace, H. S. (1977). Trial and intertrial durations in autoshaping. *Journal of Experimental Psychology-Animal Behavior Processes, 3*, 264-284.

Griffiths, O., & Le Pelley, M. E. (2009). Attentional changes in blocking are not a consequence of lateral inhibition. *Learning & Behavior, 37*, 27-41.

Griffiths, O., & Mitchell, C. J. (2008). Selective attention in human associative learning and recognition memory. *Journal of Experimental Psychology: General, 137*, 626-648.

Hawkins, R. D., Carew, T. J., & Kandel, E. R. (1986). Effects of interstimulus interval and contingency on classical conditioning of the Aplysia siphon withdrawal reflex. *Journal of Neuroscience, 6*, 1695-1701.

Hume, D. (1740/1978). *A Treatise of Human Nature*. Oxford: Oxford University Press.

Kamin, L. J. (1968). Attention-like processes in classical conditioning. In M. R. Jones (Ed.), *Miami symposium on the prediction of behavior: Aversive stimulation* (pp. 9-32). Coral Gables, FL: University of Miami Press.

Kendler, H. H., & Kendler, T. S. (1962). Vertical and horizontal processes in problem solving. *Psychological Review, 69*, 1-16.

Kendler, T. S., & Kendler, H. H. (1959). Reversal and nonreversal shifts in kindergarten children. *Journal of Experimental Psychology, 58*, 56-60.

Kendler, T. S., & Kendler, H. H. (1970). An ontogeny of optional shift behavior. *Child Development, 41*, 1-27.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22-44.

Kruschke, J. K. (2003). Attention in learning. *Current Directions in Psychological Science, 12*, 171-175.

Kuhn, D., & Pease, M. (2006). Do children and adults learn differently? *Journal of Cognition and*

Development, 7, 279-293.

Le Pelley, M. E. (2004). The role of associative history in models of associative learning: A selective review and a hybrid model. *Quarterly Journal of Experimental Psychology, 57B*, 193-243.

Le Pelley, M. E. (2010). Attention and human associative learning. In C. J. Mitchell & M. E. Le Pelley (Eds.), *Attention and Associative Learning: From Brain to Behaviour* (pp. 187-215). Oxford: Oxford University Press.

Le Pelley, M. E., Beesley, T., & Griffiths, O. (2014). Relative salience versus relative validity: Cue salience influences blocking in human associative learning. *Journal of Experimental Psychology: Animal Behavior Processes, 40*, 116-132.

Le Pelley, M. E., Beesley, T., & Suret, M. B. (2007). Blocking of human causal learning involves learned changes in stimulus processing. *Quarterly Journal of Experimental Psychology, 60,* 1468-1476.

Le Pelley, M. E., & McLaren, I. P. L. (2001). Retrospective revaluation in humans: Learning or memory? *Quarterly Journal of Experimental Psychology*.

Le Pelley, M. E., Pearson, D., Griffiths, O., & Beesley, T. (2015). When goals conflict with values: Counterproductive attentional and oculomotor capture by reward-related stimuli. *Journal of Experimental Psychology: General, 144*, 158-171.

López, F. J., Cobos, P. L., & Caño, A. (2005). Associative and causal reasoning accounts of causal induction: Symmetries and asymmetries in predictive and diagnostic inferences. *Memory & Cognition, 33*, 1388-1398.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review, 82*, 276-298.

Matute, H., Arcediano, F., & Miller, R. R. (1996). Test question modulates cue competition between causes and between effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 182-196.

McNally, G. P., & Cole, S. (2006). Opioid receptors in the midbrain periaqueductal gray regulate prediction errors during Pavlovian fear conditioning. *Behavioral Neuroscience, 120,* 313-323.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207-238.

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin, 117,* 363-386.

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

Ost, J. W. P., & Lauer, D. W. (1965). Some investigations of salivary conditioning in the dog. In W. F. Prokasy (Ed.), *Classical Conditioning* (pp. 192-207). New York: Appleton-Century-Crofts.

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

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian conditioning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review, 87,* 532-552.

Pearson, D., Donkin, C., Tran, S. C., Most, S. B., & Le Pelley, M. E. (2015). Cognitive control and counterproductive oculomotor capture by reward-related stimuli. *Visual Cognition*. Advance online publication. doi: 10.1080/13506285.2014.994252

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68,* 29-46.

Rescorla, R. A. (1967). Pavlovian conditioning and its proper control procedures. *Psychological Review, 74,* 71-&.

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology, 66,* 1-&.

Rescorla, R. A. (1972). Informational variables in Pavlovian conditioning. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 6). New York: Academic Press.

Rescorla, R. A. (1984). Signaling intertrial shocks attenuates their negative effect on conditioned suppression. *Bulletin of the Psychonomic Society, 22*, 225-228.

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 research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.

Rottman, B. M. (this volume). The acquisition and use of causal structure knowledge. In M. R. Waldmann (Ed.), *The Oxford Handbook of Causal Reasoning*. Oxford, UK: Oxford University Press.

Rozin, P., & Kalat, J. W. (1971). Specific hungers and poison avoidance as adaptive specializations of learning. *Psychological Review, 78*, 459-486.

Schlottmann, A. (1999). Seeing it happen and knowing how it works: How children understand the relation between perceptual causality and underlying mechanism. *Developmental Psychology, 35*, 303-317.

Schneiderman, N., & Gormezano, I. (1964). Conditioning of the nictitating membrane of the rabbit as a function of CS-US interval. *Journal of Comparative and Physiological Psychology, 57*, 188-195.

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

Shanks, D. R. (1986). Selective attribution and the judgment of causality. *Learning and Motivation, 17*, 311-334.

Shanks, D. R. (1989). Selectional processes in causality judgment. *Memory & Cognition, 17*, 27-34.

Shanks, D. R. (1995). *The Psychology of Associative Learning*. Cambridge, UK: Cambridge University Press.

Shanks, D. R. (2007). Associationism and cognition: Human contingency learning at 25. *Quarterly Journal of Experimental Psychology, 60*, 291-309.

Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgment. *Psychology of*

Learning and Motivation, 21, 229-261.

Shanks, D. R., & Dickinson, A. (1991). Instrumental judgment and performance under variations in action-outcome contingency and contiguity. *Memory & Cognition, 19*, 353-360.

Shanks, D. R., & López, F. J. (1996). Causal order does not affect cue selection in human associative learning. *Memory & Cognition, 24*, 511-522.

Shanks, D. R., López, F. J., Darby, R. J., & Dickinson, A. (1996). Distinguishing associative and probabilistic contrast theories of human contingency judgment. *The Psychology of Learning and Motivation, 34*, 265-311.

Tangen, J. M., & Allan, L. G. (2004). Cue interaction and judgments of causality: Contributions of causal and associative processes. *Memory & Cognition, 32*, 107-124.

Tassoni, C. J. (1995). The least mean squares network with information coding: A model of cue learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 193-204.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Review, 8*, Monograph supplement.

Van Hamme, L. J., Kao, S. F., & Wasserman, E. A. (1993). Judging interevent relations: From cause to effect and from effect to cause. *Memory & Cognition, 21*, 802-808.

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonrepresentation of compound stimulus elements. *Learning and Motivation, 25*, 127-151.

Wagner, A. R. (1981). SOP: A model of automatic memory processing in animal behaviour. 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. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology-Learning Memory and Cognition, 26*, 53-76.

Waldmann, M. R. (2001). Predictive versus diagnostic causal learning: Evidence from an overshadowing paradigm. *Psychonomic Bulletin & Review, 8*, 600-608.

Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: Two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 216-227.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General, 121*, 222-236.

Waldmann, M. R., & Holyoak, K. J. (1997). Determining whether causal order affects cue selection in human contingency learning: Comment. *Memory & Cognition, 25*, 125-134.

Wasserman, E. A., & Berglan, L. R. (1998). Backward blocking and recovery from overshadowing in human causal judgement: The role of within-compound associations. *Quarterly Journal of Experimental Psychology, 51B*, 121-138.

Wasserman, E. A., Chatlosh, D. L., & Neunaber, D. J. (1983). Perception of causal relations in humans: Factors affecting judgments of response-outcome contingencies under free-operant procedures. *Learning and Motivation, 14*, 406-432.

Table 1. Design of the within-subjects blocking experiment by McNally and Cole (2006).

Stage 1	**Stage 2**	**Test**	**Result**
A → shock	AB → shock	B	Greater fear of B than D
	CD → shock	D	

Note: A and C are visual stimuli (constant light and flashing light, counterbalanced across subjects); B and D are auditory stimuli (white noise and clicker, counterbalanced across participants).