Beyond the information (not) given: Representations of stimulus absence in rats *(Rattus norvegicus).*

Running header - Uncertainty & associations in rats

Dominic M. Dwyer (Cardiff University & University of New South Wales)

and

Michael R. Waldmann (University of Göttingen)

July 2015 – Accepted for publication in the Journal of Comparative Psychology

For correspondence: DwyerDM@cardiff.ac.uk or michael.waldmann@bio.uni-goettingen.de

Abstract

Questions regarding the nature of non-human cognition continue to be of great interest within cognitive science and biology. However, progress in characterising the relative contribution of “simple” associative and more “complex” reasoning mechanisms has been painfully slow – something that the tendency for researchers from different intellectual traditions to work separately has only exacerbated. This paper re-examines evidence that rats respond differently to the non-presentation of an event than they do if the physical location of that event is covered. One class of explanation for the sensitivity to different types of event absence is that rats' representations go beyond their immediate sensory experience and that covering creates uncertainty regarding the status of an event (thus impacting on the underlying causal model of the relationship between events). A second class of explanation, which includes associative mechanisms, assumes that rats represent only their direct sensory experience and that particular features of the covering procedures provide incidental cues that elicit the observed behaviours. We outline a set of consensus predictions from these two classes of explanation focusing on the potential importance of uncertainty about the presentation of an outcome. The example of covering the food-magazine during the extinction of appetitive conditioning is used as a test-case for the derivation of diagnostic tests that are not biased by preconceived assumptions about the nature of animal cognition.

Keywords: Causal model, renewal, secondary reinforcement, ambiguity

"And no man, when he hath lighted a lamp, covereth it with a vessel, or putteth it under a bed: But he putteth it on a stand." Luke, Ch. 8, V 16.

<u>*Putting lamps under bushels*</u>

While a lamp under a bushel casts just as little light as an unlit lamp, the status of the unlit lamp is clear, while that of the covered lamp is uncertain – it may be lit or unlit. Although probably not the typical message taken from this parable, it exemplifies the fact that, considered rationally, there is a clear difference between the absence of an event, and the absence of information about that event. One goal of the present article is to examine recent research on the capacities of rats to reason about hidden objects as a test case for examining distinctions between higher-level cognitive processes and basic associative mechanisms. But before turning our attention to these empirical concerns we will comment, relatively briefly, on the sometimes rancorous debate concerning the commonalities and differences between human and non-human animal cognition.

Comparisons between human and non-human animal cognition have attracted great interest in cognitive science and biology in the past decades. Perhaps the dominant tradition has been to assume that non-human animals are convenient systems in which to study simple processes (e.g. of learning and memory), and their underlying biological substrates, untrammelled by the more complex reasoning and rule-based processes possessed by humans. This view has been challenged by recent evidence which suggests that animals might, in addition to simple associative processes, also have far richer ways of representing the causal texture of their environment (e.g., Blaisdell, Sawa, Leising, & Waldmann, 2006; Fast & Blaisdell, 2011; Leising, Wong, Waldmann, & Blaisdell, 2008; Murphy, Mondragon, & Murphy, 2008; Waldmann, Schmid, Wong, & Blaisdell, 2012). However, the potentially far-reaching implications of these studies depend on the idea that behaviours consistent with complex cognitive mechanisms are indeed the result of such complex mechanisms, and cannot be explained as emergent properties of more simple (in particular associative) mechanisms (Burgess, Dwyer, & Honey, 2012; Dwyer, Starns, & Honey, 2009; Kutlu & Schmajuk, 2012). A fundamental shortcoming of this debate is that it is not entirely clear how higher-level cognitive processes can theoretically and empirically be distinguished from basic associative mechanisms. We present here a new proposal for making this distinction.

In the literature, different proposals have been discussed on how to distinguish higher-level cognition from associative processes. The traditional view, inspired by behaviourism, was that cognitive but not associative theories postulate information processing mechanisms operating on mental representations of the world. This distinction is no longer pertinent because many modern associative theories assume that animals possess mental representations, and characterise learning as the formation of associative links between these representations. A prime example of this is the idea that classical conditioning reflects the formation of an excitatory association between mental representations of a conditioned stimulus (CS) and an unconditioned stimulus (US) – an idea included in essentially all accounts of associative learning regardless of their differences concerning the details of the learning algorithm involved (e.g., Esber & Haselgrove, 2011; Harris, 2006; Le Pelley, 2004; Mackintosh, 1975; Pearce, 2002; Pearce & Hall, 1980; Rescorla & Wagner, 1972; Wagner, 1981). While contemporary associative theory does include (and require) mental representations, it should be recognised that these are informationally "thin" representations, held to consist essentially as copies or traces of aspects of the sensory and motivational stimulation produced by experience of the stimulus (Heyes, 2012). In particular, associative theories do not allow that either their representations or the links between them have semantic content – that is their truth value cannot be assessed. In this sense "thick" representations are effectively propositional (i.e. they can be expressed as a statement with a truth value – e.g. "The light is on" – which is either true or false, and also allows the possibility "I don't know”). In contrast, as a copy or trace of the activation produced by the stimulus, thin representations accord to nothing more than the set of nodes/elements that are activated by experience with the stimulus (or activated through associative links). Therefore, it makes no sense to ask whether the activation is “correct”, it is merely a matter of whether activation exists and to what degree. Although the fact that contemporary associative theory admits mental representations at all removes one classical divide between associative processes and complex cognition, the commitment to thin mental representations has one critical consequence: It requires associative theory to deal only with the sample of events experienced by an organism and the activation of the representations that occur as a result of this experience.

Levels of Representation

Our main focus in this article is on causal representations. Predicting and explaining events on the basis of observations and interventions is arguably one of the most important cognitive competencies that allow organisms to adapt to the world. There are a vast number of competing theories specifying the cognitive mechanisms underlying this competency. As a first approximation, we would like to propose two different classes of theories that can be distinguished on the basis of the postulated representations of the world. Of course, within each class there are numerous competing variations that have been the focus of extensive research.

Level 1: Sample-based theories:

The basic assumption underlying this class of theories is that causal representations use representations of temporally ordered observed events (cues, outcomes) and that the goal of learning is to capture the statistical relations between these events. Thus, the key assumption for our purposes is that Level 1 accounts assume that organisms do not (or cannot) look beyond the observed sample of events. The sample of learning events is what organisms know about the particular aspect of the world they observe.

One of the key topics within this class of theories is to investigate which statistical rules organisms actually use to represent the observed covariations. A large number of such rules have been proposed both within cognitive theories (e.g., Hattori & Oaksford, 2007; Perales & Shanks, 2007) and within associative theories (e.g., Dickinson, 2001; Le Pelley, Oakeshott, Wills, & McLaren, 2005; Shanks & Dickinson, 1987). One thing all these otherwise competing theories have in common is that they compute some index of covariation from the learning sample, which encapsulates the effective strength of the causal relation. Indeed, the fact that some associative and cognitive models make identical predictions under some circumstance – see for example relationship between the output of the Rescorla-Wagner model and delta-P metric discussed by Shanks (1995) – implies that these models often capture the same functional relationships between experienced events perspective (for a more detailed analysis of the implications of examining learning at a functional level see De Houwer, Barnes-Holmes, & Moors, 2013; De Houwer et al., this volume). In the present context, it is most important that such theories do not include a role for any awareness about the fallibility of experiences of the world (e.g., absence of evidence) or of the representations themselves (e.g., dreams, hallucinations vs. experiences of real events). The fact that many associative models are based around error-correction mechanisms does mean that they calculate a prediction error between the associative activation of representational nodes and the activation produced by experience of events. However, this is an algorithmic comparison and does not require the organism to have a meta-representational appreciation of the current internal associative model, the current external input, and the relationship between them. In short, sample-based theories do not assume a meta-representational understanding by the organism of the distinction between its representation of the world and the world that produces that representation.

Various research paradigms view human and non-human organisms as focusing on samples, unable to go beyond the information given. In causal research, associative theories are a prime example of this class of theories. Indeed, the fact that associative theories are characterised by a reliance on thin mental representations of stimuli and the links between them requires that they must focus on an organism's sample of experience. Thin representations do not allow an assessment of truth value, so there is no way in which the mental representation activated by a stimulus (or its activation through memory or associative means) can be evaluated as accurately corresponding to the outside world or not[1]. Moreover, thin representations ascribe no content to an associative link other than as a means for specifying the degree to which activity of one representation will influence the degree of activation in a representation to which it is associatively linked. As such associative accounts do not explicitly distinguish between causal and non-causal relationships between events.

According to this sample-based class of theories, organisms encode the presence and absence of temporally ordered events and learn statistical covariations between these events. The strength of these covarations determines inferences or behaviour. Rule-based theories of

[1] It is instructive to note here Holland's (1990) work showing that stimulus representations activated associatively ("images" in his terminology) can elicit some of the same processing that occurs when the stimulus itself is presented. The same body of work also established that the processing of retrieved images is not exactly the same as that for experienced events – so there is clearly some distinction between retrieved and directly activated stimulus representations. However, when only thin representations are assumed then this distinction in what is activated by experience (the world) and through association (the image) is literally just that, a difference in what is activated – only from the outside can the different sets of activated elements be related to which set accords to the real world. As we will see later, recent model-based accounts are very different in assuming that there is some ability to distinguish the model from the experience.

causal reasoning are another example (for a review, see, Waldmann & Hagmayer, 2013). These theories debate which exact covariation rule organisms employ. But as in the associative framework, statistical covariations are based on what is observed in a sample. In social psychology, there is also a variant of the sample view (see, Fiedler, 2012; Fiedler & Juslin, 2006). Here the claim is that judgmental biases are often caused by distortions in the observed or retrieved sample of experiences. Fiedler (2012) argues that humans are largely unable to understand and correct statistical distortions in the sample. He has labelled this deficit “metacognitive myopia.”

Level 2: Causal Models:

This class of theories assumes that organisms go beyond the information given when learning about causal relations to make inferences about an underlying unobservable causal model (see Waldmann, Hagmayer, & Blaisdell, 2006). Of course, going beyond the sample is not an all-or-none feature. There are different degrees of inferences transcending the sample, and different organisms may differ in the extent to which they are capable of going beyond the information given (for an example within causal model theory, see Waldmann, Cheng, Hagmayer, & Blaisdell, 2008).

A key difference between causal and associative theories concerns the links between causes and effects. Causal links, often depicted as arrows, are directed from cause to effect. In associative theories, temporal order determines whether an association is excitatory or inhibitory, but this alone does not result in the explicit representation that the first event caused the second. Indeed, causal and temporal order can be dissociated (e.g., Waldmann, 2000; Waldmann & Holyoak, 1992). For example, physicians often observe the symptoms (i.e., effects) prior to diagnosing the cause. The exact meaning of the causal arrows differs across theories, but the general assumption is that causal processes are unobservable and need to be inferred based on observations and prior knowledge. For example, Cheng’s (1997) power PC theory assumes that people are capable of inferring the power of a cause based on covariation and background assumptions. Power is a point estimate of the unobservable probability of the cause generating or preventing a specific effect in the hypothetical absence of background factors.

A less abstract account assumes hidden forces and causal mechanisms that transfer some kind of conserved quantity (such as linear momentum or electric charge to take examples from physics) between causes and effects (see Waldmann & Hagmayer, 2013, for a review). Although causal mechanisms can sometimes be elaborated as chains of observable variables, the variables within the chain are connected via arrows that code some kind of hidden flow of a conserved quantity (Dowe, 2000). Mechanism theories do not necessarily assume elaborate knowledge, as it is well known that human laypeople often have no or only very sketchy knowledge of the exact relationships between events (Rozenblit & Keil, 2002). The assumption rather is that people understand a relation between two events as causal if they assume that there is some kind of mechanism that links the events, even if the details of this mechanism are largely unknown.

A more recent development in causal model theory goes one step further in separating observed samples from underlying unobservable generating models. Inspired by Bayesian statistical inference, it is assumed that a rational approach to causal inference would require taking into account the fact that samples are noisy reflections of the hidden generating causal models. Thus, depending on statistically relevant factors, such as sample size, samples carry more or less *uncertainty* about the structure and the parameters of the causal model. According to this view, organisms are mainly interested in a faithful representation of the characteristics of the causal model, and therefore need to take into account uncertainty when making inferences. A number of studies have demonstrated that human subjects are indeed sensitive to statistical uncertainty (Griffiths & Tenenbaum, 2009; Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008; Meder, Mayrhofer, & Waldmann, 2014)[2].

Testing the Level of Representation

Level 1 associative and Level 2 causal model theories are often pursued in separation. A typical research strategy of those interested in either class of account is to design studies that test between competing theories within their class – while questions of between-class comparisons tend to be considered most seriously only after publication when conclusions are challenged externally. For example, it is not uncommon for alternative associative Level 1 "killjoy" (Shettleworth, 2010) accounts to be developed in a post-hoc fashion after novel patterns of behaviour had been discovered based on predictions of Level 2 theories. In this light it is rather unsurprising that progress in this area often appears meagre: if for nothing else than publication lag "conversations" in the literature are incredibly slow. In addition there is often a strong bias for Level 2 theorists to interpret data that is consistent with predictions of their complex accounts as evidence for their theory without considering the possibility that level 1 accounts of the same data might be available (this is especially prevalent when human subjects are involved). When alternative Level 1 accounts are considered, this consideration is often constrained by a lack of familiarity with contemporary associative theory. On the other hand, the emergent properties of Level 1 theories are not always apparent without considering the exact experimental situation and by themselves

[2] The nomenclature we have adopted (Level 1 vs Level 2) is entirely abstract and we admit that this may appear uninformative, but the choice was quite deliberate. While we focus here on the nature of the representations assumed at each level and the differences in terms of the explicit role of causal relationships, the distinction between these two classes of model goes beyond causality (as our subsequent discussion of theory of mind illustrates). Thus the abstract nomenclature avoids overly-restrictive characterisations of the model classes we are discussing.

Level 1 theories commonly provide little guide to the investigation of the sort of phenomena predicted by Level 2 theories. For example, it was only after Couchman, Coutinho, Beran, and Smith (2010) published their analysis of delayed feedback as supporting a (Level 2) metacognition account of primate behaviour in a discrimination task that Le Pelley (2012) was able to simulate their experimental procedures with a (Level 1) reinforcement learning account. Similarly, the demonstration that rats' behaviour can diverge as a function of whether a cue appears as a result of their actions or not followed from the prediction from a (Level 2) causal model account suggesting a critical difference between seeing and doing (Blaisdell et al., 2006). Only following the publication of the experimental methods used to produce this demonstration could Kutlu and Schmajuk (2012) examine the possibility that their associative model might be able to simulate the observed behaviour[3]. Thus, Level 1 theorists often need to await progress within Level 2 theories before they can address the question of whether the discovered phenomena genuinely require complex representations or can also be explained by a Level 1 account. One possible response to these systemic problems is the direct collaboration between researchers from different theoretical perspectives.

Of course, developing an alternative Level 1 account for a phenomenon generated by Level 2 research is only the first step. Although considerations of simplicity enshrined in Morgan's Canon (Morgan, 1894) have often led researchers, at least from the associative camp, to favour Level 1 over Level 2 theories, it should be remembered that the Canon is (at best) a guide to interpretation and does not have any logically probative status (for a more detailed discussion of this point, see Heyes, 2012). Indeed, any heuristic arguments that

[3] This far from a one-way relationship as demonstrated by the example of Bayesian reasoning accounts (e.g., Gopnik et al., 2004; Griffiths & Tenenbaum, 2009) developed to explain cue-competition effects such as backward blocking that were first reported in the associative literature.

might be applied – from considerations of parsimony to appeals to predictive or explanatory scope – cannot on their own conclusively decide between Level 1 and Level 2 accounts. As ever in science, empirical data are paramount, and thus the most productive research strategy is to develop competing Level 1 and Level 2 accounts of a phenomenon and then deploy experimental paradigms that allow differentiation between them.

But before moving to consider a test case for a targeted empirical comparison of Level 1 and Level 2 theories, we should emphasise that they are not necessarily mutually exclusive. In cognitive psychology, two-process theories (see, Evans, 2012) have become increasingly popular. One example, related to our target phenomenon, is the two-process model of theory of mind inferences by Apperly and Butterfill (2009). A typical task in this domain is the Sally scenario in which the protagonist Sally hides an object, which in her absence is transferred to a different location. The key finding is that children younger than 4 seem unable to understand that Sally will look at the place she has hidden the object regardless of the current location. When asked where she will go, young children tend to point to the actual location of the object. Fully understanding this situation requires the competency to have meta-representations that separate reality from (possibly erroneous) mental representations. Many researchers argued that young children as well as animals lack such meta-representational capacities. In the last decade, however, researchers using more implicit habituation paradigms have demonstrated some level of understanding of this task even in infants (Onishi & Baillargeon, 2005). Apperly and Butterfill therefore postulate two separate processes that may underlie the responses in the different tasks. Whereas infants may only understand that agents look for something where they have seen it last, older children may reason with more complex meta-representations, which in the beginning stages of reasoning leads to the observed errors. According to the two-process view, some species may only be capable of reasoning with the simpler process, whereas others may have both types of processes at their disposal. Critically however, even for these sort of two-process accounts, the question remains as to whether a particular behaviour is (or can be) supported by the simpler process or only the more complex one. So the importance of determining the representational level at which an organism is functioning remains germane even from the perspective of dual-process accounts.

Hidden Events: A Simple Test Case for Sensitivity to Uncertainty

The present article will discuss a fairly simple potential indicator of uncertainty, uncertainty about the status of events. Level 2 causal model accounts would differentiate between two possible causes for the failure to experience an expected event: Either the event is really absent in the world, or the event is present but access to it is being prevented in some fashion. Waldmann et al. (2012) examined a test-case for this possibility in the extinction of Pavlovian appetitive conditioning. In their experiments, rats were presented with three learning and test phases. In Phase 1, an association between a cue (CS), a light, and sucrose (US) was established through a Pavlovian conditioning procedure (a 10s light was presented and the offset of the light followed by 10s access to a sucrose-filled dipper)[4]. In Phase 2, the extinction phase, the cue was paired with the experience of absence of sucrose (the light was presented in advance of the empty dipper – i.e. the dipper arm was raised for 10s, but the trough did not contain sucrose, so no primary reward was presented). Then in Phase 3, the degree of extinction was tested by presenting the light cue without sucrose (again, the empty dipper continued to be presented). The crucial manipulation involved Phase 2. In one

[4] The food magazine was positioned above a trough containing sucrose solution. A mechanical dipper arm, with a small cup on the end, was immersed in this solution. Sucrose access was provided by raising the arm so that the cup protruded through a hole in the base of the food magazine for 10s before being lowered again. The rats could not access either the dipper arm or the sucrose except when it was raised.

condition, the No-Cover condition, rats could directly observe that sucrose was actually absent from the food magazine, whereas in the alternative Cover condition a metallic plate was placed over the magazine preventing rats from accessing it. The test phase showed that rats differentiated between these conditions with greater test phase responding to the CS in the Cover than the No-Cover condition. Moreover, it was not merely the presence of the metallic plate that controlled responding, because a control condition where the plate was included without preventing access to the food magazine did not prevent extinction.

As noted above, the causal model account would interpret this finding as evidence that rats are capable of differentiating between two possible causes of the absence of sucrose in the extinction phase: Either the sucrose is really absent, or it is present but access is blocked. This inference requires an understanding of uncertainty of the status of events. In other words, initial training experience should create a *light causes sucrose* model. The transition from the rewarded training phase to the non-rewarded extinction phase could potentially create an ambiguity in a causal understanding of the situation – has the causal relationship changed, and the light no longer causes sucrose to appear, or is the relationship still is intact but the sucrose has for some other reason not been observed? This ambiguity would be emphasised when access to the usual source of sucrose delivery was prevented during extinction – although the light is still experienced without sucrose, both possible causal structures are still consistent with the experience because there is no direct disconfirmation of the expected sucrose delivery. Thus a causal model analysis would suggest that covering the sucrose magazine should attenuate the effects of extinction and help preserve the *light causes sucrose* model. In turn, preserving a causal relationship between the light and sucrose should result in higher responding in the test phase - which is exactly what happened (Waldmann et al., 2012). Clearly, a full causal understanding of this situation requires some kind of understanding of the difference between the representations of the world and the actual world. Even in humans, unless people have philosophical training, this differentiation is unlikely to be explicitly available. It suffices that in specific cases absence is distinguished from lack of evidence.

Functionally the separation between experience and world has a number of potential advantages for organisms. If experience and the world were collapsed, every instance of disappearance due to another object blocking sight would lead to a fading of the representation of the object although it is still present behind the occluder. Since such experiences are common, the physical representation of the world arising from such inferences would be very different from ours. Work on object permanence with animals seems to indicate that many animals may not think that objects behind an occluder actually disappear from the world (Gómez, 2004, 2005). Similarly, in Waldmann et al.'s (2012) study organisms that only represent present and absent events and do not differentiate between absence in the world and lack of evidence would represent events in Phase 2 (extinction) as a gradual change of contingency. Although this is certainly a possibility, as the No-Cover condition demonstrates, it is not necessarily adaptive to always make this inference. One key feature of causal relations is that they tend to be stable and do not suddenly change (Pearl, 2000). Thus, the capacity to distinguish between different causes of experienced absence is potentially adaptive for an organism that has the goal of forming veridical representations of the causal texture of the world and if these veridical representations improve the organism's success in interacting with the world.

Associative Accounts of Hidden Events: Renewal and Secondary Reinforcers

As described above, a causal model account based on uncertainty can explain why covering the food magazine during extinction might result in higher levels of responding during test. However, the details of the experiments performed also admit alternative explanations of the same results based entirely on associative Level 1 mechanisms: We will consider one based on response prevention[5], a second based on renewal theory, and another on a consideration of conditioned reinforcement.

Rescorla (2001) notes that there is typically a direct relationship between the amount of non-reinforced responding in extinction and the degree to which such non-reinforcement impacts on future behaviour. For example, following tone-food pairings, presentation of the tone alone will typically result in some degree of responding to the food magazine during an extinction phase, while devaluation of the food reward or satiating the animals reduces the level of extinction phase magazine responding. Even though the number of unrewarded tone alone presentations is unaffected by devaluation or satiation, these treatments which reduce extinction phase magazine responding also reduce the effectiveness of extinction (Holland & Rescorla, 1975). On the basis of such results, Rescorla (2001; see also Colwill, 1991) suggested that learning not to make a particular response may make a critical contribution to the decrement in responding typically observed in extinction. One direct corollary of this idea is that the effects of non-reward in extinction will be reduced if the original response is not produced. In the present circumstances, covering the magazine clearly prevents the target response of magazine entry, and thus prevention of this response should protect it from extinction. Not only does this provide a simple explanation of why test phase responding was be higher after the magazine was covered in the extinction phase, it also explains why introducing a similar metallic cover that did not prevent access to the magazine had little effect.

A second associative account of the effects of the magazine cover comes from renewal theory. This approach suggests that extinction should be specific to the context in

[5] We would thank one of the reviewers of an earlier version of this paper for their suggestion of this possibility.

which it occurs, and that extinguished responses should reappear when testing occurs in a situation more akin to the original training context than to the context of extinction (e.g., Bouton, 2004; Delamater, 2004). In the current situation, the cover provided during extinction could act as a context change, so its removal would comprise a return to the original training context, thus supporting the re-emergence of responding. Thus, according to this view rats would gradually start to represent Phase 2 as a situation in which the light is paired with the absence of sucrose, but expression of this new association would be restricted to the context in which extinction took place. This possibility was acknowledged in the original report of these experiments, and in Experiment 3 of that paper an additional control group was used in which the metal "cover" was inserted into the apparatus during the extinction phase, but did not actually prevent access to the food magazine. This control, in which the presence or absence of a cover could have acted as a cue separating the extinction and text contexts, resulted in performance that was no different to that in the No-Cover condition. However, it may be argued that a cover preventing access to a source of food is more salient than a cover placed elsewhere, in which case a magazine cover would be a more effective contextual cue than one that does not cover the magazine.

It should be noted that in all the Cover conditions the sucrose dipper continued to be raised and lowered, but that there was "no noticeable vibrations for the human ear" (p. 983, Waldmann et al., 2012), that could be discerned inside the experimental chamber. That is, covering was assumed to have prevented all access to information about the operation of the dipper during extinction[6]. Thus in the covering situation, the training and test contexts were

[6] It should be noted that this assumption was not directly tested, and given that rat and human sensory abilities are somewhat different then it is certainly plausible that the rats in Waldmann et al.'s (2012) experiments were able to sense some aspect(s) of the dipper's operation behind the cover. Although this possibility has no direct impact on the ideas discussed here, it does raise the issue of what predictions the different accounts of the

similar in the operation of the dipper but diverged from the extinction context in both respects – while in the No-Cover, and the plate without covering conditions, the extinction and test contexts both included the operation of an empty dipper. In short, covering the magazine in the extinction phase of the experiments produced several potential cues that could have differentiated the extinction and test contexts. This could support the recovery of extinguished responding in the covered condition without reference to any Level 2 mechanisms.

The final alternative account of the covering data we will consider here relies on secondary reinforcement. Remembering that the training phase of these experiments was based on pairing the light with a sucrose filled dipper, the training phase should establish light-sucrose, light-dipper, and dipper-sucrose associations. It is well known that animals will respond both to cues paired with primary reinforcers - i.e. the sucrose in these studies - and also secondary reinforcers - i.e. any stimulus that is associated with a primary reinforcer (for reviews see, Mackintosh, 1974; Mackintosh, 1983). In these studies the dipper would have accrued secondary reinforcing properties by being paired with sucrose during the training phase. Following this, all groups received light-alone presentations in the extinction phase - presumably extinguishing light-sucrose associations to a similar extent between groups. In the No-Cover condition the empty dipper would also be experienced – resulting in the extinction of the dipper-sucrose associations, and thus the removal of secondary reinforcing properties of the dipper. However, in the Cover condition, the dipper would not be experienced at all during the extinction phase, which would protect the dipper-sucrose associations and preserve the conditioned reinforcement properties of the dipper. In turn, this would allow the dipper to support responding to the light when the light was again paired

---

covering effect might make regarding “partial” covers (e.g. explicitly preventing vision but not audition).

with the dipper in the test phase. In short, the training phase paired the light cue with both a primary (sucrose) and a secondary (the sucrose-paired dipper) reinforcer. Covering the magazine in the extinction phase of the experiments could preserve the secondary reward properties of the dipper compared to the uncovered conditions. The secondary reinforcing properties of the dipper could support additional test-phase responding in the covered condition without reference to any Level 2 mechanisms.

Divergent predictions from Level 1 and Level 2 accounts of hidden events

One important feature of the causal uncertainty and renewal/secondary reinforcement accounts of the effects of covering the magazine is that the differences between them relate directly to the nature of the division between Level 1 and Level 2 theories outlined previously. The causal model account suggests that uncertainty produced by the cover would preserve the strength of a *light causes sucrose* model in the face of experiencing the light without sucrose. This goes beyond the direct sample of experience because the fact that sucrose did not follow the light is discounted due to a distinction between absence of sucrose (the No-Cover case) and absence of evidence (the Cover case). That is, the effects seen in the test phase are a product of covering producing uncertainty over whether the sucrose did or did not occur, and thus reducing the effective level of extinction. In contrast, the three associative accounts considered here all related to direct effects of the cover in extinction or its removal at test. The response-prevention account suggests that covering reduces the effects of extinction because the target response could never be produced when the magazine was covered. Both the renewal and secondary reinforcement accounts assume that extinction does occur due to experience of the light without sucrose, but that responding returns in the test phase due to events that happen during that test: For renewal theory, the critical event in the Cover condition is that the context of test is different from that of extinction (it allows access to the magazine and includes an operating dipper – as in training but not extinction); For secondary reinforcement, the critical event is that the rats experience the light paired with the dipper, and in the Cover condition the dipper will be a secondary reinforcer (but not in the No-Cover condition, because then the previous experience of the empty dipper has removed the secondary reinforcing properties of the dipper) – these test phase light-dipper pairings support the re-acquisition of responding to the light. That is, the associative accounts are sample-based as they refer only to events that are actually experienced (or not experienced, in the case of prevented responses). Therefore, empirical tests of the divergence between these accounts speak not only to the particular details of each of them, but also to the more general division between Level 1 and Level 2 processes in the context of this behavioural procedure[7].

Effects of manipulating dipper presentation:

Given that the status of the dipper in the extinction and test phases is critical to two of the Level 1 sample-based accounts, while uncertainty concerning the presence of reward is central to the Level 2 causal model account, one empirical test would be to manipulate the presence of the dipper during these phases. That is, to compare the pattern of responses between groups that receive either: (A) training and testing as in the original paper with the empty dipper presented during the extinction and test phases; or (B) with no presentation of the empty dipper during either the extinction or test phases (i.e. the dipper would remain lowered – but not be explicitly removed from the chamber). Table 1 outlines the proposed experiment and summarises the key predictions of each of the accounts for responding to the light at the beginning of the test phase of the experiment. The original experiments included control conditions which received extinction without the magazine cover. Such controls are

[7] Of course, it is also possible to assess how causal models might account for the direct effects of test phase events, but this would not address our current concern with whether rats are able to go beyond the sample of their experience in terms of the explicit role for uncertainty.

needed to establish a baseline for levels of responding after effective experimental extinction, and we would propose including such uncovered controls which would receive extinction and test with or without dipper presentation in the current experiment. Although it is likely that the operation vs. non-operation of the dipper would influence the rate of experimental extinction, we will not considered these control conditions in any detail because (as in the original experiments) the extinction phase would be continued until responding to the light has stopped, and so all theoretical accounts would predict negligible test phase responding. The derivation of the predictions for the critical magazine cover conditions is fleshed out in turn for the causal model, response prevention, renewal, and secondary reinforcement accounts.

In both the Dipper Cover and No-Dipper Cover conditions the training phase would produce a *light causes sucrose* model. In the extinction phase, the light occurs alone, but because access to the magazine is blocked the *light causes sucrose* model will be protected because the covering means that the status of the sucrose is uncertain and thus the evidence for sucrose not appearing is partially or totally discounted in terms of relevance to the light-sucrose relationship. Covering might also protect the light-sucrose causal relationship because it leads to the formation of a more complex causal model whereby the light causes sucrose but the action of an external event stops this being expressed (e.g. the cover stops access to the delivered sucrose). In the test phase, the cover is removed – so behaviour will be determined by the *light causes sucrose* model (i.e. moderate to high responding is predicted). Critically, the extinction phases for the Dipper Cover and No-Dipper Cover conditions are the same. In both conditions, the dipper and sucrose are covered during extinction so the causal model at the start of test should be the same. In turn, this same causal model predicts that the response to the light at the start of test would be the same in these two conditions. Of course, as the test phase continues, then the Dipper Cover and No-Dipper

Cover conditions will have different experiences. Thus their causal models, and levels of responding, may be expected to diverge across testing: for example, the non-operation of the dipper might support the formation of a more complex causal model whereby the light causes sucrose only through the action of the dipper, which for some reason did not operate (e.g. the dipper was stuck). However, the dipper is operated at the end of the light during training, so at the time of responding is assessed (during the presentation of the light) there is no direct evidence to indicate whether or not the dipper will operate on that trial. So even if responding is dependent on the expectation of dipper operation, this expectation should only decline gradually as the light is encountered without the dipper following immediately afterwards. Irrespective of these issues, responding early in the test phase should remain diagnostic of the strength of the light-sucrose causal relationship at the end of the extinction phase to the extent that causal representations are stable (Pearl, 2000).

The predictions of the response-prevention account are simple – in both the No-Dipper Cover and Dipper Cover conditions the cover will prevent the production of magazine entry responses. To the extent that extinction requires the production of the relevant response, then such response prevention will attenuate the effects of extinction, and levels of magazine responding to the light would be predicted to be high at the start of the test phase.

As outlined above, the renewal account suggests that the training phase should establish an excitatory light-sucrose association, while presenting the light without the reward in extinction will create an inhibitory light-“no sucrose” association. Responding at test will be determined by the degree to which these two associations are expressed – something that is controlled by the similarity of the extinction and test phase contexts. For the Dipper Cover condition, the test phase and the extinction phase differ in two critical respects, access to the magazine and the operation of the dipper: both of which are absent in the extinction phase and present at test. Thus, the extinction and test contexts are quite different which will attenuate the expression of the inhibitory light-“no sucrose” association formed in extinction and result in responding to the light on the basis of the originally-formed excitatory light-sucrose association – a classic renewal effect. In contrast, for the No-Dipper Cover condition, the test phase and the extinction phase differ with respect to access to the magazine, but are the same with respect to the non-operation of the dipper. Thus, while there will be some difference between the extinction and test contexts in the No-Dipper Cover condition, and thus some degree of renewal would be expected, this should not be as great as in the Dipper Cover condition. As the non-operation of the dipper can only be observed after the first trial, this difference between the Dipper and No-Dipper conditions should emerge across the extinction phase.

Finally, the conditioned reinforcement account is based on the potential contribution of the dipper as a secondary reinforcer due to its pairing with sucrose in the training phase of the study. In the Dipper Cover condition, the light is presented in the absence of either the primary or secondary reinforcer during the extinction phase – so by the end of extinction there will be no effective source of primary or secondary reinforcement. However, the secondary reinforcing properties of the dipper will be preserved through the extinction phase because the dipper is never experienced without sucrose. In the test phase, the light will again be presented in conjunction with the dipper, and thus the secondary reinforcing properties of the dipper will support responding to the light (at least for as long as the dipper remains an effective secondary reinforcer). Obviously, this secondary reinforcing effect of the dipper could only be apparent after the first trial of the extinction phase. The No-Dipper Cover condition will also result in the removal of any effective source of primary or secondary reinforcement by the end of the extinction phase, but in this case dipper operation is not reintroduced at the test phase. So test phase responding to the light will be low in this condition.

In summary, all accounts predict that, if the dipper continues to be presented, then covering the magazine in extinction will result in higher levels of test phase responding than if the magazine is uncovered in extinction. Two of the associative accounts – renewal and secondary reinforcement – predict that this covering effect will be reduced or removed if the dipper is not presented after the training phase. In contrast, uncertainty within a causal model account and the response prevention account both predict that the effects of covering the magazine will be preserved, at least in the initial trials of the test phase in which the absence of the dipper is not yet apparent.

Importantly, these predictions emphasise the test phase as a whole. However it has already been noted that the presence or absence of the dipper might produce changes in the levels of responding across the test phase. We have not considered trial-by-trial effects in the predictions we have described thus far. The predictions of associative theories regarding changes during extinction depend on the assumed learning parameters. Cognitive theories would predict that changes of expectation depend on prior knowledge about causal stability within the learning domain (e.g., physical vs. social). Little is known about these effects. However, the very first trial of the test phase is different from all subsequent trials because the response to the light is assessed before the dipper is presented (or not presented) and so the Dipper versus No-Dipper manipulation cannot influence responding on the first test trial. The impact of this fact is particularly clear in terms of the secondary reinforcement account as it predicts that responding should emerge after only after the light is followed by the dipper. Similarly, the renewal account predicts some responding to the light on the first trial in the Dipper Cover and No-Dipper Cover conditions (because the removal of the cover is a return to part of the training context), but only after the first trial will the Dipper vs No-Dipper manipulation contribute to the context change between extinction and test phases. Therefore, it should be recognised that the theoretical accounts we have presented here do imply that responding could vary in a systematic fashion across trials, and that the different accounts make divergent predictions about such trial-by-trial effects. That said, it should also be acknowledged that the variability in responding that motivates the usual practice of aggregating across multiple trials may make a reliable assessment of such fine-grained predictions difficult in practice.

Sign-tracking vs. Goal-tracking:

Thus far, we have discussed responding to the light, following light-sucrose pairings, entirely in terms of a single measure – magazine entry. However, Pavlovian conditioning can establish a range of possible responses when a cue stimulus is paired with reward (Boakes, 1977). In particular, a distinction is made between sign-tracking, i.e., responding directed towards the conditioned stimulus, and goal-tracking, i.e., responding towards the unconditioned stimulus (for recent examples of this distinction in the context of cues predicting food reward, see Flagel, Watson, Robinson, & Akil, 2007; Meyer et al., 2012). In the present context, the original light to sucrose training should establish both a sign-tracking response (e.g. orientation to the light) and a goal-tracking response (e.g. entry to the sucrose magazine). Clearly, covering the sucrose magazine in extinction will prevent animals from producing the same goal-tracking responses they produced in the training phase, but would have no impact on the production of sign-tracking responses to the light. Therefore, an examination of sign-tracking and goal-tracking responses would shed some light on the mechanisms underpinning the effects of covering the food magazine during extinction. On a practical note, sign-tracking to a light can be assessed by videoing the animals and measuring the number of times the orient to the light. However, many studies of sign- vs goal-tracking have used a retractable lever as the CS (Flagel et al., 2007; Meyer et al., 2012). Here, a lever is inserted and removed from the box just as a light may be turned on and off. Critically, the lever is entirely a signal; there is no need for the rats to press it in order for the reward to be delivered. Despite this, rats will still approach and press the lever, and thus sign-tracking can be measured by the number of lever presses, while goal tracking can continue to be assessed through magazine entry. Table 2 outlines a proposed experiment using these techniques and summarises the key predictions of each of the accounts in terms of sign and goal tracking responses. This experiment would use a lever as the cue in place of the light used in previous experiments to facilitate recording of sign-tracking responses, but all other aspects of the experiment would remain the same. That is, the critical condition involves covering the food magazine in the extinction phase. We will focus our analysis on this condition although a control group receiving extinction without the magazine would still be needed to establish the effects of experimental extinction for comparison purposes. As before, the derivation of these predictions is fleshed out in turn for the causal model, response prevention, renewal, and secondary reinforcement accounts.

The predictions of the causal model approach are based on the uncertainty surrounding the appropriate causal structure. However, cognitive theories have not as yet addressed how exactly expectations translate into different types of behaviour. Because the relationship between model-based expectation and behavioural measures have not been the subject of detailed consideration we have assumed here that, for all responses, a simple monotonic function relates the degree of expectation of reward to the level of response[8]. Critically rats that are sign-tracking respond towards to a cue to the extent that it reliably predicts reward, and rats that are goal-tracking respond to the site of reward delivery during the presentation of the cue, again, to the extent that the cue reliably predicts rewards. Thus both sign- and goal-tracking behaviours are determined by the cue to reward relationship. In terms of the causal model account described here this reflects the strength of the *light causes*

[8] This represents a minimal assumption which allows the causal model approach to reflect the fact that both goal- and sign-tracking behaviours occur. It also focuses our analysis only on the effects of uncertainty regarding sucrose presentation in the extinction phase.

sucrose model. As described above, this model might be protected from the effects of extinction through the creation of uncertainty about the status of the reward by covering of the magazine. Under these preliminary assumptions, the consideration of uncertainty within the causal model account predicts that both sign- and goal-tracking responses will be affected by covering the sucrose magazine during the extinction phase.

As noted above, covering the magazine will prevent goal tracking (i.e. magazine entry) responses, but would not prevent sign-training (i.e. lever press) responses. To the extent that extinction requires the production of the relevant response, then covering the magazine will attenuate the effects of extinction on goal-tracking responses but will not influence the extinction of sign-tracking responses. Therefore, the action of response prevention alone predict that levels of magazine responding to the light would be high at the start of the test phase, while levels of lever press responding would be low.

With respect to the renewal account, the local context for the goal-tracking response is the magazine. Covering the magazine is a distinct and salient change to this local context and so the covering manipulation will mean that magazine responses at test will occur in a different context to that experienced during extinction. As described above, this difference in context between extinction and test phases should produce renewal and thus levels of magazine responding (i.e. the goal tracking response) would be expected to be high at test. In contrast, the local context for the sign-tracking response is the lever, which is not directly affected by the covering manipulation. Thus, although the global context will differ between extinction and test due to the presence/absence of the magazine cover, the local context for sign-tracking responding will be the same for extinction and test. This similarity in the local context for extinction and test should act to support generalisation of learning in extinction to the test phase. Thus, while some renewal is expected for sign-tracking responses, this will less than that seen for goal-tracking, and so renewal theory predicts that levels of lever-press responding at test would be moderate.

The predictions of the secondary reinforcement account are somewhat less categorical. Both sign- and goal-tracking after covering should relate to the same CS-US relationship – where the effective US here is the conditioned reinforcement provided by the dipper. So if covering preserved the conditioned reinforcing properties of the dipper then both sign- and goal-tracking responses should return after the dipper is paired with the light during test. However, there are large individual differences between animals in the levels of sign- and goal-tracking responses they produce (Flagel et al., 2007; Meyer et al., 2012), and animals that display a preponderance of sign-tracking responses may have a reduced opportunity to interact with the conditioned reinforcer during the test phase. If so, then the conditioned reinforcement account also predicts a greater effect of the covering manipulation on goal-tracking than sign-tracking responses.

In summary, how uncertainty is translated into sign- and goal-tracking behaviours has not been specified yet within the class of theories which includes causal model approaches. Under the preliminary assumption that all responses reflect the strength of the underlying *light causes sucrose* model, the causal model account predicts that sign- and goal-tracking responses will both be affected by the magazine covering manipulation because uncertainty about the status of the sucrose reward will protect this causal model. The three Level 1 associative accounts all relate to direct effects of the covering manipulation through either preventing only one of the target responses in extinction, having different effects on the local context for lever press and magazine entry responses, or by influencing the interaction with the secondary reward. Thus the response competition and renewal accounts (and to a less certain extent the secondary reinforcement account), predict that goal-tracking responses should be more sensitive to magazine covering in extinction than sign-tracking responses.

Summary and comparisons to previous approaches

In the initial parts of this paper we outlined a distinction between two general classes of theoretical accounts: Level 1 – which refers to accounts that focus on the representations of events as experienced by the organism, and (in associative versions of such account at least) involve only thin, non-semantic representations of events and the links between them; and Level 2 – which refers to accounts that are focused on the idea that sensory experience is the basis for forming models of the events in the world and the nature of the relationships between them (with a particular focus on causal relationships), and thus involve explicitly semantic representations of events. We then considered one test case involving extinction of a classically conditioned CS-US relationship, where covering the food magazine during the extinction phase attenuated the effects of that extinction in a subsequent test. While both Level 1 and Level 2 accounts of the observed behaviour are available, these accounts make divergent predictions about the effects of manipulating the details of how the reward was delivered and the nature of the response assessed. Critically, these divergent predictions speak directly to the level at which the theoretical accounts were based: The Level 1 accounts are based only on sensitivity to manipulations influencing the precise events experienced by the animals in the test phase; while the Level 2 account we have considered is focused on how covering the magazine creates uncertainty regarding the presence or absence of the reward, which in turn will impact on how experiencing the absence of sucrose modifies the causal model of the situation that was established during initial training. This influence of uncertainty on the *light causes sucrose* model is explicitly a level 2 account as it clearly goes beyond the direct effects of the sample of events experienced.

It should, of course, be noted that while the predictions of the four accounts (uncertainty in causal models, response prevention, renewal, and secondary reinforcement) are clear, it would be entirely possible to make post-hoc revisions or additions to them. For example, a renewal theorist may suggest that the key feature of the context was not the dipper but some other aspect of the magazine. Moreover, it should be emphasised that we have focused the causal model account entirely on the effects that covering might have by inducing animals to go beyond the direct effects of experience through creating uncertainty. But all causal theories, regardless of their sensitivity to uncertainty, also assume Level 1 contingency learning competencies. For example, on a causal account one could assume that the dipper is part of the causal model learned in the acquisition phase (light-dipper-sucrose) so that its absence in the test phase would lead to changes of expectation. These changes would be solely due to Level 1 causal contingency learning which should be unaffected by the cover manipulation in the extinction phase. That said, the current experiments do make a direct comparison between an explanation in terms of uncertainty alone (i.e. an example of a Level 2 "beyond the sample" account) and explanations in terms of particular local features of the manipulations (i.e. examples of Level 1 "sample-based" accounts). Thus, while the two experimental manipulations described here do not comprise a definitive and general test of causal model theory and its associative alternatives on their own, they do provide a specific test of whether uncertainty over the presence or absence of reward considered alone is able to explain the behaviour of animals in the current extinction situation.

We think it is instructive to compare our current approach – based on directly examining one key (Level 2) aspect of a causal model account – with previous approaches. In addition to the extinction experiments considered here, there are several other demonstrations that preventing rats having access to the source of significant stimulus events results in behaviour that is materially different to the simple non-presentation of those events (Blaisdell, Leising, Stahlman, & Waldmann, 2009; Fast & Blaisdell, 2011). These other covering experiments were discussed by Dwyer and Burgess (2011), but only to present

Level 1 associative accounts of the observed behaviours and to dismiss the originally-proposed Level 2 accounts entirely on the basis of an appeal to Morgan's Canon. That is, there was no discussion of how to make an empirically-based comparison between the alternative accounts let alone any report of new or relevant empirical data. So, while the Dwyer and Burgess analysis was of value in providing an existence-proof of an associative account, it makes no progress towards determining whether the behaviour of the rats was under the control of Level 1 or Level 2 mechanisms.

In summary, the current paper attempts to approach the investigation of the cognitive mechanisms underpinning the behaviour of human and non-humans animals without bias from preconceived assumptions regarding the prior probability of one account over another. This approach supported the derivation of diagnostic empirical tests focusing on the key feature of the current situation (i.e. the effect of uncertainty) which divided the current theoretical accounts on the basis of the general level of representation they instantiate. Of course, the proof of this particular pudding is in the baking, and we are in the process of preparing to run exactly the studies we outline here.

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116*(4), 953-970.

Blaisdell, A. P., Leising, K. J., Stahlman, W., & Waldmann, M. R. (2009). Rats distinguish between absence of events and lack of information in sensory preconditioning. *International Journal of Comparative Psychology, 22*(1), 1-18.

Blaisdell, A. P., Sawa, K., Leising, K. J., & Waldmann, M. R. (2006). Causal reasoning in rats. *Science, 311*(5763), 1020-1022.

Boakes, R. A. (1977). Performance on learning to associate a stimulus with positive reinforcement. In H. Davis & H. M. B. Hurwitz (Eds.), *Operant Pavlovian interactions* (pp. 67-97). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learning & Memory, 11*(5), 485-494.

Burgess, K. V., Dwyer, D. M., & Honey, R. C. (2012). Re-assessing causal accounts of learnt behavior in rats. *Journal of Experimental Psychology: Animal Behavior Processes, 38*(2), 148-156.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*, 367-405.

Colwill, R. M. (1991). Negative discriminative stimuli provide information about the identity of omitted response-contingent outcomes. *Animal Learning & Behavior, 19*(4), 326-336.

Couchman, J. J., Coutinho, M. V. C., Beran, M. J., & Smith, J. D. (2010). Beyond Stimulus Cues and Reinforcement Signals A New Approach to Animal Metacognition. *Journal of Comparative Psychology, 124*(4), 356-368.

De Houwer, J., Barnes-Holmes, D., & Moors, A. (2013). What is learning? On the nature and merits of a functional definition of learning. *Psychonomic Bulletin & Review, 20*(4), 631-642.

Delamater, A. R. (2004). Experimental extinction in Pavlovian conditioning: Behavioural and neuroscience perspectives. *Quarterly Journal of Experimental Psychology Section B-Comparative and Physiological Psychology, 57*(2), 97-132.

Dickinson, A. (2001). Causal learning: An associative analysis. *Quarterly Journal of Experimental Psychology, 54B*(1), 3-25.

Dowe, P. (2000). *Physical Causation*. Cambridge: Cambridge University Press.

Dwyer, D. M., & Burgess, K. V. (2011). Rational accounts of animal behaviour? Lessons from C. Lloyd Morgan's canon. *International Journal of Comparative Psychology, 24*, 349-364.

Dwyer, D. M., Starns, J., & Honey, R. C. (2009). "Causal Reasoning" in rats: A reappraisal. *Journal of Experimental Psychology: Animal Behavior Processes, 35*(4), 578-586.

Esber, G. R., & Haselgrove, M. (2011). Reconciling the influence of predictiveness and uncertainty on stimulus salience: a model of attention in associative learning. *Proceedings of the Royal Society B-Biological Sciences, 278*(1718), 2553-2561.

Evans, J. S. B. T. (2012). Dual-process theories of reasoning: Facts and fallacies. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 115-133). New York, NY: Oxford University Press; US.

Fast, C. D., & Blaisdell, A. P. (2011). Rats are sensitive to ambiguity. *Psychonomic Bulletin & Review, 18*(6), 1230-1237.

Fiedler, K. (2012). Meta-cognitive myopia and the dilemmas of inductive-statistical inference. In B. H. Ross (Ed.), *Psychology of Learning and Motivation, Vol 57* (Vol. 57, pp. 1-55).

Fiedler, K., & Juslin, P. (2006). *Information sampling and adaptive cognition*. New York, NY: Cambridge University Press; US.

Flagel, S. B., Watson, S. J., Robinson, T. E., & Akil, H. (2007). Individual differences in the propensity to approach signals vs goals promote different adaptations in the dopamine system of rats. *Psychopharmacology*, *191*(3), 599-607.

Gómez, J. C. (2004). *Apes, monkeys, children and the growth of mind*. Cambridge, MA: Harvard University Press.

Gómez, J. C. (2005). Species comparative studies and cognitive development. *Trends in Cognitive Sciences, 9*(3), 118-125.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111*(1), 3-32.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review, 116*(4), 661-716.

Harris, J. A. (2006). Elemental representations of stimuli in associative learning. *Psychological Review, 113*(3), 584-605.

Hattori, M., & Oaksford, M. (2007). Adaptive non-interventional heuristics for covariation detection in causal induction: Model comparison and rational analysis. *Cognitive Science, 31*(5), 765-814.

Heyes, C. (2012). Simple minds: a qualified defence of associative learning. *Philosophical Transactions of the Royal Society B-Biological Sciences*, *367*(1603), 2695-2703.

Holland, P. C. (1990). Event representation in Pavlovian conditioning: Image and action. *Cognition*, *37*, 105-131.

Holland, P. C., & Rescorla, R. A. (1975). The effect of two ways of devaluing the unconditioned stimulus after first- and second-order appetitve conditioning. *Journal of Experimental Psychology-Animal Behavior Processes, 1*(4), 355-363.

Kutlu, M. G., & Schmajuk, N. A. (2012). Classical conditioning mechanisms can differentiate between seeing and doing in rats. *Journal of experimental psychology. Animal behavior processes, 38*(1), 84-101.

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 Section B-Comparative and Physiological Psychology, 57*(3), 193-243.

Le Pelley, M. E. (2012). Metacognitive monkeys or associative animals? Simple reinforcement learning explains uncertainty in nonhuman animals. *Journal of Experimental Psychology-Learning Memory and Cognition, 38*(3), 686-708.

Le Pelley, M. E., Oakeshott, S. M., Wills, A. J., & McLaren, I. P. L. (2005). The outcome specificity of learned predictiveness effects: Parallels between human causal learning and animal conditioning. *Journal of Experimental Psychology-Animal Behavior Processes, 31*(2), 226-236.

Leising, K. J., Wong, J., Waldmann, M. R., & Blaisdell, A. P. (2008). The special status of actions in causal reasoning in rats. *Journal of Experimental Psychology-General, 137*(3), 514-527.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review, 115*(4), 955-984.

Mackintosh, N. J. (1974). *The psychology of animal learning*. Oxford, England: Academic Press.

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

Mackintosh, N. J. (1983). *Conditioning and associative learning*. Oxford: Carendon Press.

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review*, *121*(3), 277-301.

Meyer, P. J., Lovic, V., Saunders, B. T., Yager, L. M., Flagel, S. B., Morrow, J. D., & Robinson, T. E. (2012). Quantifying individual variation in the propensity to attribute incentive salience to reward cues. *PLoS ONE, 7*(6).

Morgan, C. L. (1894). *An introduction to comparative psychology*. London, England: Walter Scott Publishing Co; England.

Murphy, R. A., Mondragon, E., & Murphy, V. A. (2008). Rule learning by rats. *Science*, *319*(5871), 1849-1851.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*(5719), 255-258.

Pearce, J. M. (2002). Evaluation and development of a connectionist theory of configural learning. *Animal Learning & Behavior*, *30*(2), 73-95.

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

Pearl, J. (2000). *Causality*. Cambridge, UK: Cambridge University Press.

Perales, J. C., & Shanks, D. R. (2007). Models of covariation-based causal judgment: A review and synthesis. *Psychonomic Bulletin & Review, 14*(4), 577-596.

Rescorla, R. A. (2001). Experimental extinction. In R. R. Mowrer & S. B. Klein (Eds.), *Handbook of contemporary learning theories* (pp. 119-154). Mahwah, NJ: Lawrence Erlbaum.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F.

Prokasy (Eds.), *Classical Conditioning II: Current theory and research.* (pp. 64-99). New York: Appelton-Century-Crofts.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science, 26*(5), 521-562.

Shanks, D. R. (1995). Is human learning rational? *Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology, 48A*(2), 257-279.

Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgment. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. Vol. 21, pp. pp. 229-261). San Diego, CA, US: Academic Press, Inc.

Shettleworth, S. J. (2010). Clever animals and killjoy explanations in comparative psychology. *Trends in Cognitive Sciences, 14*(11), 477-481.

Wagner, A. R. (1981). SOP. A model of autonomic memory processing in animal behavior. In N. E. Spear & 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*(1), 53-76.

Waldmann, M. R., Cheng, P. W., Hagmayer, Y., & Blaisdell, A. P. (2008). Causal learning in rats and humans: a minimal rational model. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind. Prospects for Bayesian Cognitive Science* (pp. 453-484). Oxford: Oxford University Press.

Waldmann, M. R., & Hagmayer, Y. (2013). *Causal reasoning*. New York, NY: Oxford University Press; US.

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*(6), 307-311.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal-models: Asymmetries in cue competition. *Journal of Experimental Psychology-General, 121*(2), 222-236.

Waldmann, M. R., Schmid, M., Wong, J., & Blaisdell, A. P. (2012). Rats distinguish between absence of events and lack of evidence in contingency learning. *Animal Cognition, 15*(5), 979-990.

Table 1 – Dipper Manipulation

Condition	Train	Extinction	Test	Uncertainty & Causal Model	Response Prevention	Renewal	Secondary Reinforcement
Dipper Cover	Light to sucrose filled dipper	Light alone & dipper magazine covered	Light to empty dipper	Status of reward uncertain in extinction phase – this protects *light causes sucrose* model. Expression of causal model at test supports responding to light. I.e. Test phase responding moderate to high (depending on degree of protection by uncertainty).	Cover prevents magazine response, therefore extinction effect of light alone presentation reduced for this response. I.e. Test phase responding high.	Extinction and test phases differ in presence of the cover and dipper operation. This is a large difference between extinction and test phases, so expect renewal. I.e. Test phase responding high.	Primary reward (sucrose) removed. Secondary reward properties of dipper preserved as the dipper is not experienced without sucrose in extinction. Secondary reward can support responding at test. I.e. Test phase responding moderate.
No-Dipper Cover	Light to sucrose filled dipper	Light alone & dipper magazine covered	Light alone			Extinction and test phases differ with in presence of cover, but are the same in the non-operation of the dipper. This is a smaller difference between extinction and test phases than in the Dipper Cover condition. So expect some renewal, but not as much as in Dipper Cover condition. I.e. Test phase responding moderate.	Primary reward (sucrose) and secondary (dipper) removed. Neither primary nor secondary reward can support responding at test. I.e. Test phase responding low.

Note 1: These predictions assume the cover completely blocks all access to the operation of the dipper. As an operational means to ensure this assumption is accurate, in the both the Dipper Cover, and No-Dipper Cover conditions, the dipper would not be operated at all in the extinction phase.
Note 2: Cells have been merged to highlight where predictions are not affected by the key manipulation.
Note 3: Additional control conditions where the extinction phase takes place without a magazine cover (e.g. Dipper No-Cover and No-Dipper No-Cover) would be needed in order to establish the baseline level of responding, these have not been illustrated here as all accounts predict experimental extinction and negligible responding at test.

Table 2 – Sign- vs Goal-tracking

Condition	Train	Extinction	Test	Uncertainty & Causal Model	Response Prevention	Renewal	Secondary Reinforcement
Dipper Cover Measure sign-tracking (lever press)	Lever insertion to sucrose filled dipper	Lever alone & Dipper magazine covered	Lever to empty dipper	Status of reward uncertain in extinction phase – this protects *light causes sucrose* model. Expression of causal model at test supports responding. I.e. Test phase responding moderate to high for lever and magazine entry (depending on degree of protection by uncertainty).	Cover does not prevent lever response, therefore extinction from lever alone presentation expected. I.e. Test phase lever responding low.	Local context for sign tracking response is lever, which is unchanged between extinction and test phase. Unchanged local context attenuates renewal effect based on global context change due to extinction and test phases differing in presence of the cover and dipper operation. I.e. Test phase responding to the lever moderate.	Primary reward (sucrose) removed. Secondary reward properties of dipper protected by covering but high levels of orienting to lever may reduce experience of dipper as secondary reward. Secondary reward can support responding at test to the extent it is experienced. I.e. Test phase responding to the lever moderate to low.
Dipper Cover Measure goal-tracking (magazine response)	Lever insertion to sucrose filled dipper	Lever alone & Dipper magazine covered	Lever to empty dipper	Status of reward uncertain in extinction phase – this protects *light causes sucrose* model. Expression of causal model at test supports responding. I.e. Test phase responding moderate to high for lever and magazine entry (depending on degree of protection by uncertainty).	Cover prevents magazine response, therefore extinction effect of lever alone presentation reduced for this response. I.e. Test phase magazine responding high.	Local context for goal tracking response is the magazine. Extinction and test phases differences (magazine cover and dipper operation) focused on magazine. This is a large difference between extinction and test phases so expect renewal. I.e. Test phase magazine responding high.	Primary reward (sucrose) removed. Secondary reward properties of dipper protected by covering. Secondary reward can support responding at test. I.e. Test phase magazine responding moderate.

Note1: This is a within-subject experiment with sign- and goal-tracking responses measured in all animals – however, the panels have been split to illustrate where different predictions are made for different response types.
Note 2: As with the previous experiment, these predictions assume the cover completely blocks all access to the operation of the dipper. As an operational means to ensure this assumption is accurate, in the Dipper Cover condition, the dipper would not be operated at all in the extinction phase.
Note 3: Again, additional control conditions where the extinction phase takes place without a magazine cover would be needed in order to establish the baseline level of responding, these have not been illustrated here as all accounts predict experimental extinction and negligible sign or goal tracking responding at test.