# Can Rats Reason?

Stephane Savanah
Macquarie University

Since at least the mid-1980s claims have been made for rationality in rats. For example, that rats are capable of *inferential reasoning* (Blaisdell, Sawa, Leising, & Waldmann, 2006; Bunsey & Eichenbaum, 1996), or that they can make *adaptive decisions about future behavior* (Foote & Crystal, 2007), or that they are capable of knowledge in *propositional-like form* (Dickinson, 1985). The stakes are rather high, because these capacities imply concept possession and on some views (e.g., Rödl, 2007; Savanah, 2012) rationality indicates self-consciousness. I evaluate the case for rat rationality by analyzing 5 key research paradigms: *spatial navigation*, *metacognition*, *transitive inference*, *causal reasoning*, and *goal orientation*. I conclude that the observed behaviors need not imply rationality by the subjects. Rather, the behavior can be accounted for by noncognitive processes such as hard-wired species typical predispositions or associative learning or (nonconceptual) affordance detection. These mechanisms do not *necessarily* require or implicate the capacity for rationality. As such there is as yet insufficient evidence that rats can reason. I end by proposing the 'Staircase Test,' an experiment designed to provide convincing evidence of rationality in rats.

*Keywords:* causal reasoning, goal orientation, metacognition, rationality, transitive inference

Recently researchers have become excited about the possibility that rats may be rational (e.g., Clayton & Dickinson, 2006). The issue is of great import because rationality[1] is a high-level cognitive ability with major implications for our view of animals. For example, Savanah (2012) argues that the concept possession required for reasoning implies self-consciousness and Rödl (2007) argues that a nexus exists between self-consciousness and reasoning, which he even claims is the "principle thought of Kant and Hegel" (p. 109). We still debate whether even our closest animal relatives, chimpanzees,

are self-conscious, so a conclusion that rats are self-conscious is of major significance. Howsoever, even without accepting a link between rationality and self-consciousness we should insist on a high level of evidence before accepting that rats can reason.

In what follows I examine experimental paradigms in which researchers claim rationality in rats explicitly or implicitly, as indicated by the use of terms such as 'inference,' 'problem-solving,'[2] 'propositional thought,' and so on. I consider examples of five key and arguably representative research paradigms and allied experiments to examine in this regard: spatial navigation, metacognition, transitive inference, causal reasoning, and goal orientation. I conclude that the evidence for rationality in rats is not yet convincing. I end by proposing an experimental paradigm (the Staircase Test) that sets the bar very high for demonstrating ratio-

---

[1] I elaborate further on my usage of 'rationality' under "What Counts as Rationality?" but I wish it to be clear from the start that by rationality I mean explicitly the ability to reason.

[2] Where 'problem-solving' refers specifically to a process of inferential reasoning rather than coming across a solution via random actions.

nality yet would provide convincing evidence that rats can reason.

## What Counts as Rationality?

The concept of rationality I investigate refers to the process of conscious, inferential reasoning, or what Kacelnik (2006) refers to as 'PP-rationality' (for 'philosophical/psychological' usage): "To judge whether behavior is PP-rational one needs to establish if it is caused by beliefs that have emerged from a reasoning process" (p. 89). Familiar examples of reasoning in humans are deductive reasoning, inductive reasoning and abductive reasoning, the antithesis of which is behavior that is elicited as an automatic response. Of course it is a fair assumption that cognitive capacities fall along a continuum between these putative extremes. Indeed, with regard to observations of animal behavior in general and laboratory rats in particular there is growing acceptance that the "available comparative evidence does not fit comfortably into either the traditional associationist or classically inferential alternatives" (Penn & Povinelli, 2007, p. 98). Furthermore, it is possible that rational and nonrational processes might coexist in the rat, as they sometimes do in humans. Indeed, under "Goal Orientation" below I discuss an experiment (the 'Palermo Protocol') that may in fact indicate an intermediate condition in the rat subjects. As such, setting the threshold for what we should count as reasoning is difficult and may appear somewhat arbitrary. Keeping such considerations in mind my approach is to set the bar high, such that finding rat subjects reaching this bar would represent a substantial breakthrough. Thus, rationality as here discussed involves thoughts occurring within the 'space of reasons,' implying *concept possession*. Therefore, to accept behavior as evidence of reasoning in rats we must eliminate plausible noncognitive explanations. Accordingly, we must reject as evidence any behavior that is species typical or acquired through associative learning or is otherwise 'programmatic' and thus not reliant on possession of concepts.

A definition of 'concept' is notoriously difficult to pin down but for our purposes everyday usage will suffice. To have a concept of something is to have an idea or understanding about what it is. Concepts are required for reasoning but some entities that *appear* to behave rationally do not possess concepts. For example, chess-playing computers seem to reason but merely execute programmed instructions. They do not possess concepts and it is fair to say that most philosophers would not want to ascribe beliefs to them (Schwitzgebel, 2011). Analogous behavior in living organisms include genetically hard-wired species typical behavior or behavior shaped by Pavlovian learning, although the 'programming' in the former case occurs via natural selection and in the latter case by associative learning. Nevertheless, these programmed behaviors can be expressed by organisms that are not rational in the sense we are discussing, that is, signifying concept possession.

Of course by the above reasoning we could observe much human behavior and similarly argue that it is not evidence of concept possession. Indeed, the human brain has also been sculpted by evolution and people often behave in programmatic ways. For example, nonconscious mechanisms can influence decision-making, such as when ethnic background music in a wine shop unknowingly engenders a bias toward purchasing wines from that region (North, Hargreaves, & McKendrick, 1999). However, in the human case concept possession can be established by several means. First, we know what it is like to possess concepts—we have the personal experience of concept possession. Second, we observe in humans decidedly nonprogrammatic behaviors, such as when reasoning is applied to solve novel problems. Third, language provides direct evidence of abstract (conceptual) thought.[3] Accordingly, although it may be true that "neuroscience tells [us that] the brain is an association machine— one way or another, all of its acquired and innate functions are based on associations" (H. Eichenbaum, personal communication, June 25, 2013), it is also nevertheless true that there is something special about the human mind.

---

[3] Some would go as far as suggesting that reasoning is not even possible without natural language (Carruthers, 1996). However, I am open to the possibility that such thoughts might take place in a type of 'Mentalese' (Fodor, 1975) and consequently rats and other animals should not be precluded from the possibility of inferential thinking solely on the basis of their lack of natural language. Conceivably, nonlinguistic organisms might be able to engage in inferential thinking without the capacity to possess those thoughts in propositional form.

*Emerging* from the development of the most sophisticated of all 'association machines' is the further capacity for reasoning that transcends programmatic behavior. Studies (e.g., Frank, O'Reilly, & Curran, 2006) suggest the operation of dual mechanisms for human decision making: some on the basis of explicit reasoning processes and others on the basis of implicit reward associations. We know humans operate under this dual psychology, but do other animals? Do chimpanzees? Maybe. What about rats?

Kacelnik contrasts PP-rationality with E-rationality (economic rationality) and B-rationality (evolutionary biological rationality). E-Rationality focuses on whether behavior is consistent with the 'maximization of utility,' interpretable as maximization of energy efficiency in the context of ecology, whereas B-rationality represents behavior driven by genetic predispositions. Neither E-rationality nor B-rationality is necessarily dependent on a process of reasoning.[4] Nevertheless, in the cases I discuss there are clear explicit or implicit references to PP-rationality, implying actual inferential reasoning by the rat subjects.

The point made earlier about computer algorithms—that their behavior is programmatic and does not involve reasoning—provides one class of objection against claims of rationality in rats. What happens in the rat's brain is undoubtedly more sophisticated than a chess computer's number crunching, but if we can use computers to simulate animal responses to stimuli using noncognitive systems such as mathematical models then this suggests that the animal behavior, too, might be programmatic (e.g., De Lillo, Floreano, & Antinucci, 2001). In other words, like the computer, the behavior might not be caused by "beliefs that have emerged from a reasoning process." Of course that does not preclude rationality because rats might conceivably also behave under a dual psychology[5] like humans. But because a plausible, noncognitive alternative account exists, we cannot claim the behavior as conclusive evidence of reasoning.

Another class of objection is related to what I call 'nonzero effects.' For example, reasoning such as 'this lever provides a sickness-inducing drink, therefore I must avoid it' predicts zero lever presses but the results are usually only a relative reduction. Furthermore, a comparison with similar human results need not imply that rats are therefore capable of reasoning like humans; an alternative interpretation is that humans are capable of reacting through noncognitive processes like rats. Indeed, this last point can be generalized: where animal and human behavior in an experiment are comparable this should not necessarily be taken as strong evidence of reasoning in the rat; the evidence must be evaluated on a case-by-case basis.

I argue below that the threshold for rationality, as set earlier, has not been reached in any of the experimental paradigms discussed. In many cases the argument centers on the availability of nonrationality based accounts. I acknowledge that this does not on its own preclude the possibility of rationality in rats. Nevertheless, the claim that rats are rational is of such magnitude that the onus of proof is clearly on the claimant and we should hold such claims to a high standard of evidence. If behaviors can be accounted for without assuming rationality then those behaviors cannot be considered sufficient evidence of rationality.

## Rationality and Self-Consciousness

It is important to know whether rats or other animals are able to reason, as this would indicate concept possession. I have elsewhere argued that concept possession alone is sufficient evidence for self-consciousness (Savanah, 2012), where 'self-consciousness' is construed as *an understanding of one's own existence as a psychological subject with intentional agency*. I argue that to possess any concept the *self*-concept must be present. The argument is rooted in the theory of nonconceptual content and defends the claim that the only factor sep-

---

[4] In some cases 'maximization of utility' might involve inferential reasoning but it can also be achieved by noncognitive processes. This needs to be examined on a case-by-case basis. The point is that only PP-rationality refers specifically to the process of reasoning.

[5] Daw, Niv, and Dayan (2005) have shown computational competition exists between two areas of rat brains and that habit-formation is a process of transferring control from the prefrontal cortex to the dorsolateral striatum; however, this need not indicate a 'dual psychology' in the sense used for humans (despite the homologous brain areas in rats and humans). Rather, I suggest it is another case of E-rationality whereby processing is optimized via a Bayesian process (which indeed might apply to humans also). These issues are further discussed under "Causal Reasoning."

arating organisms that are conscious but *not* self-conscious from organisms that are both conscious *and* self-conscious is concept possession. If this claim is correct, insofar as reasoning requires concept possession, rational organisms must also be self-conscious. Thus, where convincing evidence exists for rationality in animals, including rats, I am bound to consider them self-conscious.

In an earlier work (Savanah, 2012) I emphasized that concept possession—like rationality—is most likely not an 'all-or-nothing' capacity. Indeed, neither is self-consciousness: to the extent that a creature may have only a partial capacity for concept possession, it would have partial capacity for self-consciousness. Once again, the threshold is not well-defined. Nevertheless, the higher we set the required standard of evidence the closer we come to human-like self-consciousness.

## Spatial Navigation

Even relatively simple organisms can perform impressive-seeming feats of spatial navigation. For example, Tarsitano and Jackson (1997) showed that jumping spiders of the genus Portia can navigate a route to a perceived prey even when the route includes a section forcing the spider to lose visual contact with the prey. Nevertheless, virtually all organisms, including those not usually associated with possessing conceptual abilities, navigate their environments. Accordingly, spatial navigation does not seem to rely on conceptualization or reasoning ability. Nevertheless, some authors do appear to ascribe rationality to animals based on their spatial navigation abilities (e.g., Eichenbaum, 2000).

The hippocampus in rats has long been known to play a significant role in spatial navigation (Moser, Kropff, & Moser, 2008). However, Eichenbaum (2000) has suggested that "the hippocampus may be required for new *problem solving* in familiar environments" (p. 45, emphasis added). As I discuss in later sections, we must remain cautious with respect to interpretations regarding brain regions homologous between rats and humans. Although the hippocampus probably performs some common functions across species, we do not have enough justification to infer that humans' higher hippocampal functions (such as, perhaps, its con-

tribution to reasoning) is replicated in the rats'. Indeed, evidence suggests that there is no single 'logic' module and human brain systems for reasoning are dynamically extended beyond the hippocampus (Goel, Makale, & Grafman, 2004).

Eichenbaum describes experiments using the Morris Water Maze in which a tank of water has a single column standing just below the surface as a refuge for rats: "when rats with hippocampal damage that have successfully learned to locate the escape platform from a single start position are tested from new start positions, they fail to readily locate the platform. In contrast, normal animals swim directly to the escape locus on each new probe trial" (Eichenbaum, 2000, p. 45). These results and other findings (e.g., Nadel & MacDonald, 1980) support the view that the rat hippocampus is involved in spatial memory. Nevertheless, these findings do not necessarily implicate the hippocampus in inferential thinking or any type of conceptual thought, as some seem to suggest. For example, O'Keefe and Nadel (1978) suggest that the rat's cognitive mapping system allows the animal to link conceptually diverse parts of an environment. However, their theory need not assume an advanced cognitive capacity to effectively account for rat behavior. Despite their descriptions of an animal's spatial abilities as 'knowledge' or 'knowing that,' their view easily conforms to a model devoid of conceptual abilities. Their model still works when we substitute 'information' (by which I mean to imply its nonconceptual nature) for 'knowledge' (which is conceptual) as in this example regarding the role of evolution: "an innate spatial mode of perception could be developed which would confer upon the perceiver accurate knowledge [information] of certain aspects of the external world" (p. 23). My point is that it is sufficient to consider the cognitive map as encoding *nonconceptual* spatial information that is available to hard-wired (i.e., programmatic) information processing mechanisms.

Cheng (1986) describes the rat's cognitive map as a 'metric frame' that specifies locations primarily by their geometric relations to environmental shape. The experimental evidence implicates the hippocampus in the formation and maintenance of the metric frame, but it does not follow that in using the metric frame for navigation, as in the water maze experiment,

that inferential reasoning is needed to 'solve' the problem of locating the escape platform. The metric frame might simply be the encoding of the localized environmental geometry accessible via the rat's hippocampus.

## Metacognition

Tests for metacognition in animals have mostly relied on a 'bail-out' paradigm (e.g., Foote & Crystal, 2007; Hampton, 2001, 2005; Smith, 2005; Smith, Shields, & Washburn, 2003). For example, Smith (2005) investigated the ability of rhesus macaques to judge their own confidence at visual density discrimination tasks. After training, the subjects are to judge whether a box of pixels on a computer screen is dense or sparse according to a set threshold, and are rewarded for a correct answer. For a lesser but guaranteed reward the subject also has the option to decline the test, in effect to answer with a response of 'uncertain.' Judgments of this type are deemed to be acts of metacognition because the subject bases a decision not on the density of the box but apparently on knowledge of its own ability to succeed at the task. In Hampton's (2001) paradigm the subjects must decide (after a varying delay) whether they can still remember which of four randomly selected images they had just seen. Hampton found that the longer the delay, the more likely the monkeys were to bail out of the test and settle for a lesser reward.[6]

It can be argued that metacognition construed as the capacity to monitor one's ability to discriminate between stimuli or remember an image represents a case of rationality. However, Mitchell (2002) has suggested that the monkey's choices in the Hampton (2001) experiments might only demonstrate simple rule-following without awareness of possession of the relevant mental image: if an internal image persists when the choice is presented then take the test, or if it does not persist, then bail out. Le Pelley (2012) was able to reproduce the monkey results using simulations based on a simple model of reinforcement learning. Because the simulation was based on a noncognitive system it is plausible that the monkeys also applied a noncognitive system rather than reasoning. Here I take a different approach, suggesting that the findings of experiments using the bail-out paradigm on rats can be explained in terms of first-order stimulus-response associative learning.[7]

Smith (2005) dismisses associative learning as an explanation, but his position relies on assuming that the subjects' associations to reward are dipolar (e.g., based on options such as 'dense vs. sparse' or 'same vs. different'). Below, in my analysis of the rat metacognition experiments, I argue that there is a potential for an association to be formed for a third (unintended) option—that is, the bail-out option itself represents a third stimulus that becomes associated with the lesser reward.

### Metacognition in Rats

Foote and Crystal (2007) trained rats over seven weeks at two hours per session for 35 sessions using tones of eight different durations ranging from 2 to 8 seconds. The four shorter tones were associated with a 'Left' lever for a reward, whereas the four longer ones were associated with the 'Right' lever for the reward. The two tones in the middle of the range were the most difficult to classify as long or short. In the 'Choice' condition of the test phase, after presentation of a tone, the rat had the option of entering one of two apertures. In aperture 1 ('take-the-test') were the two levers, Left and Right, in which pressing the correct lever produced a reward of six food pellets. In aperture 2 ('decline-the-test') the rat obtained a guaranteed but lesser reward of three food pellets (see Figure 1). In the 'Forced' condition only aperture 1 was available. Choice and Forced tests were intermixed randomly throughout the session. Each session lasted 9 hours and rats underwent an average of 1,546 trials.

In Choice trials, for tones near each extreme (long or short) rats were more likely to enter

---

[6] Hampton (2001) and Smith et al. (2003) observed that the choice patterns are similar between humans and monkeys in these experiments. However, as mentioned in the introduction and also noted by several other authors (e.g., see peer commentary on the Smith et al., 2003 target article by Campos & Karmiloff-Smith, 2003; King, 2003; Metcalfe, 2003; Shettleworth & Sutton, 2003; and Zentall, 2003) this is not strong evidence that the underlying mechanisms must be the same, or indeed that metacognition was necessarily involved even in the human case.

[7] Carruthers' (2003, 2008) account of the Smith (2005) results is not dissimilar to the one I present here, but there is a crucial difference: Carruthers is willing to ascribe beliefs to the subjects as part of his explanation. Holding propositional attitudes such as belief would imply concept possession, which I deny in my account.
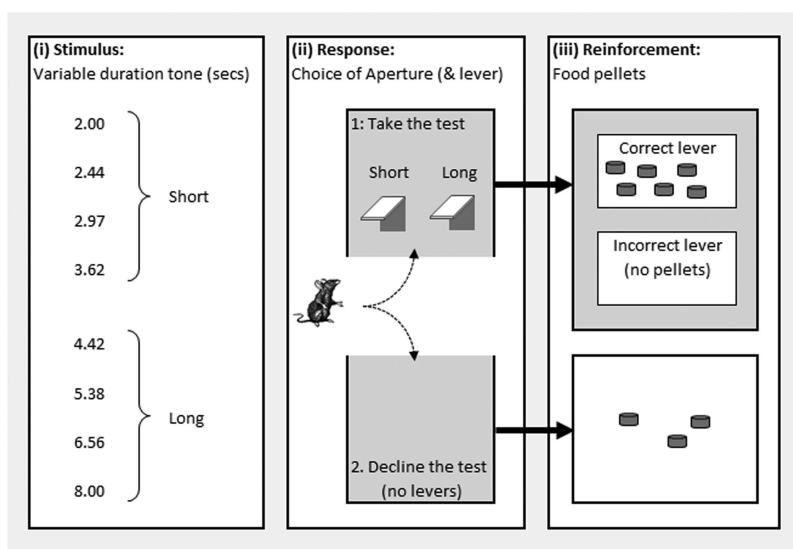
*Figure 1.* Schematic depiction of rat metacognition experiment.

aperture 1, presumably indicating confidence of their own ability to make a correct lever selection. For the harder to discriminate middle tones, the rats were more likely to enter aperture 2, presumably indicating lower confidence and so taking the option for the lower guaranteed reward. Furthermore, for the middle tones only, accuracy in the Left/Right choice was higher in the Choice trials compared with the Forced trials, which seems to support the view that the rats declined the choice because of their inability to decide on the correct lever.

Assuming that strength of association is a function of reward value and reward reliability, a stimulus–response that is rewarded every time will result in a stronger association than for one that is rewarded only half the time. Thus, a tone duration at either extreme (long or short) which the subject can easily discriminate reliably results in a high value reward, resulting in a strong association with the required response (correct lever selection). However, the middle duration tones are harder to discriminate between long or short and result in more errors being made. In those cases they are just as likely to get a reward as to *not* get a reward no matter which lever they press; therefore, it is dubious that those durations will form associations with *either* lever. So, during the training phase, although strong associations will be formed for

both the very long and very short tones, a third possibility exists that little or no association is formed for the middle tones.

Given the lack of associations formed for the middle tones during training, there is reason to believe associations formed during the test phase. In effect training did not end with the initial training phase, but continued through the relatively long test phase when the reward structure was different. Notice that in the test phase there are *three* possibilities in the reward schedule: six pellets; three pellets and no pellets. During the test phase, the rat is able to form new associations involving aperture 2, which *always* yields a small reward. These latter associations, being of lesser reward, are not likely to override the previously established stronger associations for long/short, provided, as is the case, those associations remain well rewarded. However, new associations can be established for the intermediate duration tones because they did not form strong associations during the training phase. So, along with the strong associations of the long and short tones with their paired levers and food rewards, a new association is free to be established between the middle tones and aperture 2. Instead of a choice of either 'long' or 'short' plus a metacognitive understanding of 'I cannot tell,' the choices may have actually been 'long,' 'short' and 'intermediate,' with no meta-

cognition involved. Including 'intermediate' as a possible discrimination, Table 1 lists all the associations possible according to the available behavior options.

Table 1 shows that for Choice trials there are three behavior paths for each of the three available discriminations: *long*, *short*, or *intermediate.* The final column in the table indicates the strength of incentive for each of the available actions following discrimination. For a discrimination of 'long,' the strongest incentive is behavior L3 (enter aperture 1 and select the right lever). Similarly, for a discrimination of 'short' behavior S4 is preferred (enter aperture 1 and select the *left* lever). A discrimination of 'intermediate,' I5, though weakly incentivized compared to L3 and S4, is nevertheless more favorable than the available alternatives. This associative model correctly predicts the observed behavior of the rats. As such, there is no need to assume metacognitive abilities in rats to account for these experimental results.

On the intermediate stimuli (only), the rats scored a higher proportion of correct selections on the Choice trials compared with the Forced trials. Foote and Crystal (2007) claim that this finding supports metacognitive explanations but not associative ones. In fact the three-stimulus paradigm presented above does predict this effect[8] and therefore provides more support for the noncognitive account. Consider Figure 2, which shows the stimulus duration lengths (vertical lines) along with a possible[9] range of misperceptions (error spreads) centered on each of the two intermediate stimuli (3.62 sec and 4.42 sec). When the rat's perception (or misperception) falls within the intermediate range it will enter the associated aperture for the three-pellet reward, as proposed above. Nevertheless, if it makes a larger error prompting it to 'take-the-test,' it is much more likely to err in the 'correct' direction (i.e., in the 4.42 sec case it will likely err on the 'long' side, whereas in the 3.62 sec case it will err on the short side). Therefore, on the Choice trials, the 'take-the-test' aperture will be overrepresented with fortuitous errors (inadvertently correct choices), which inflates the accuracy. This sampling bias does not apply to the Forced trials, which will have equal balance between fortuitous and unfortunate errors.

Furthermore, Foote and Crystal (2007) argue that associative accounts presuppose that, contrary to other evidence, rats are risk averse; that

is, that they prefer guaranteed small rewards over uncertain large ones. Yet quite to the contrary, noncognitive accounts presuppose no such cognitive abilities at all. Even using terminology such as 'risk-averse' and 'risk-prone' already presupposes a higher level of cognition than is necessary. We cannot tell that the rats are deliberately *judging risks*; we can only observe that they are behaving in a way that maximizes pay-off. As mentioned in the introduction, this type of behavior might only indicate a case of E-Rationality, as discussed next.

## Behavioral Economic Model (BEM)

Jozefowiez, Staddon, and Cerutti's (2009) Behavioral Economic Model explains the Foote and Crystal's (2007) results in terms of pay-off maximization—as a case of E-Rationality, which as earlier discussed is not a case of inferential reasoning. This mathematical model is based on only two assumptions: (a) when confronted with a stimulus a subject emits the behavior associated with the higher pay-off; and (b) the perception of the stimulus is noisy (i.e., the rats' perception has an error spread as discussed in the previous section). The 'noisiness' is effected with a Gaussian distribution function to simulate the spread of possible stimuli in the variable tone duration experiment. On adjusting certain parameters, such as the degree of 'preference' for lesser rewards (i.e., what Foote and Crystal might erroneously refer to as the level of 'risk aversion') assumed for the subject and the width of the Gaussian probability curve, the model does a fair job of predicting the behavior observed in the rat experiments. Accordingly Jozefowiez et al. argue that metacognition is not needed to explain the observed results in rats:

> BEM, which lacks any metacognitive ability—only basic discrimination processes—satisfies the two generally accepted criteria for metacognition: that the probability of picking the uncertain response increases with the difficulty of the task. . .and that the subject is more accurate on free-choice trials then on forced-choice trials. (p. 33)

---

[8] I am indebted to an anonymous reviewer for suggesting this analysis.

[9] The argument holds no matter how wide the range.

Table 1
*The Full Range of Possible Associations in the Foote and Crystal (2007) Experiment*

| Tone duration | Cond. | ID | Behavior | | Reward level | Trained effect |
|---|---|---|---|---|---|---|
| | | | Aperture | Lever | | |
| Long | Forced | L1 | Take test | Right | High | Strong association |
| | | L2 | Take test | Left | Zero | Negative association |
| | Choice | L3 | Take test | Right | High | Strong association |
| | | L4 | Take test | Left | Zero | Negative association |
| | | L5 | Decline | N/A | Small | Weak association |
| Intermediate | Forced | I1 | Take test | Right | (Indeterminate) | (Indeterminate) |
| | | I2 | Take test | Left | (Indeterminate) | (Indeterminate) |
| | Choice | I3 | Take test | Right | (Indeterminate) | (Indeterminate) |
| | | I4 | Take test | Left | (Indeterminate) | (Indeterminate) |
| | | I5 | Decline | N/A | Small | Weak association |
| Short | Forced | S1 | Take test | Right | Zero | Negative association |
| | | S2 | Take test | Left | High | Strong association |
| | Choice | S3 | Take test | Right | Zero | Negative association |
| | | S4 | Take test | Left | High | Strong association |
| | | S5 | Decline | N/A | Small | Weak association |

*Note.* When presented with each of the three discriminable options, this matrix correctly predicts the observed behavior.

## Transitive Inference

The phenomenon of mediated conditioning has been known for several decades (e.g., Savastano & Miller, 1998). Two different pairings containing one common element elicit an association between the two unpaired elements (e.g., (a) A paired with B; and (b) B paired with C; causes (c) A associated with C). That such cross-associations can be trained in animals is no cause—on their own—to suggest the presence of inferential reasoning. However, some researchers (e.g., Eichenbaum, 2000) have suggested that rats are capable of transitive inference, in which they *infer* specific relations between the unpaired elements. For example,

elements A and B are paired such that A is rewarded but not B, thereby training the subject to prefer A over B. Then B is paired with C such that B is rewarded but not C, so that B is preferred to C (see Figure 3). Then the subjects are presented with A and C to determine which is preferred.

Transitive inference is implicated if A is preferred over C. Note that in this paradigm it could be argued that subjects will prefer A over C because C has *never* been rewarded and this explanation does not rely on transitive inference (Allen, 2006). For this reason it is common practice to use five elements (A, B, C, D, and E) and to compare B and D so as to ensure that the
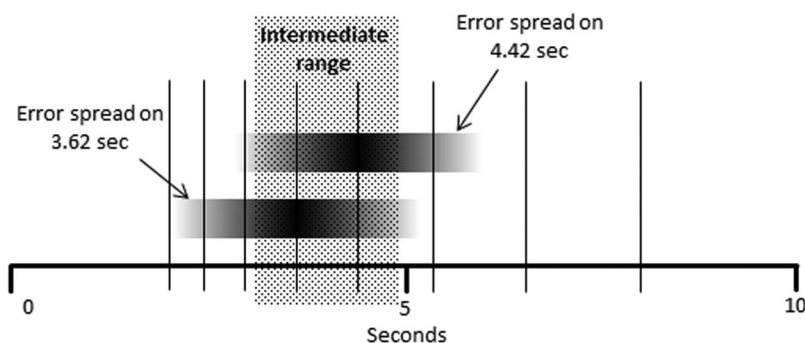


*Figure 2.* Most errors for 'intermediate' stimuli are fortuitously correct, which accounts for the greater accuracy on free-choice tests.
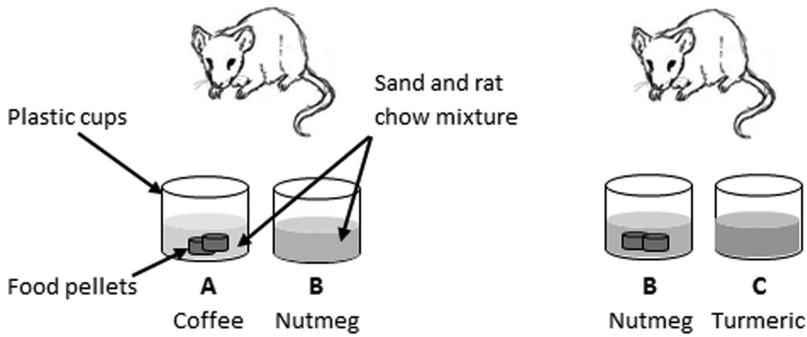
*Figure 3.* Schematic depiction of Transitive Inference experiment (three elements).

compared elements have had equal measures of reinforcement and to avoid any other boundary effects (e.g., Bunsey & Eichenbaum, 1996; Dusek & Eichenbaum, 1997). Below, I describe the experimental paradigm, an explanation of the findings that assumes *rationality* on the part of the subjects, and alternative interpretations that do *not* rely on reasoning ability.

## The Five-Element Transitive Inference Paradigm

Labeling the odors as A, B, C, D, and E, and using the '>' symbol to represent preference, the trained preferences can be expressed as A > B; B > C; C > D; D > E (Eichenbaum, 2000). In probe trials following the training, rats showed a preference of B over D (B > D) despite the fact that both were equally rewarded in their original pairings. According to Eichenbaum (2000), the rats "showed a robust capacity for transitive inference, indicating that rats are capable of linking information about the odors acquired across distinct experiences, and of making *inferential judgments based on knowledge about the orderly series*" (p. 46, emphasis added). The strength of Eichenbaum's claim that this amounts to a case of inference rests on two key factors. First, B and D are not merely associated; rather, the inherent *relation* between B and D appears to have been inferred based on the hierarchy established during training. Second, although normal rats were able to express this relation, a second group of rats with induced hippocampal damage was not, even though this latter group was able to acquire the individual pairings at normal rate. As argued next, although these findings show that the hip-

pocampus is critical to the transitive inference *effect* observed in rats, it does not necessarily support a conclusion of inferential *reasoning* in rats.

## Hippocampus in Rats and Humans

Given the importance of the hippocampus in humans for higher cognitive capacities such as declarative memory, it is tempting to ascribe the same or at least similar functionality for the hippocampus in animals like the rat. Bunsey and Eichenbaum (1996) compared their transitive inference studies in rats to analogous studies with humans and write: "in both humans and animals, stimuli can be associated independently of hippocampal function but the establishment of representations that can be expressed indirectly and inferentially is critically supported by the hippocampus" (p. 257). However, we should be cautious about ascribing higher cognitive capacities such as inferential reasoning to animals based on the similarity of human/animal brain components. Indeed, there is evidence that the hippocampus in rats is responsible for the transitive inference effect as a *noncognitive* process.

Van Elzakker, O'Reilly, and Rudy (2003) investigated the role of the hippocampus in transitive inference experiments. The researchers reproduced the five-element experimental results and also conducted a six-element version (A > B; B > C; C > D; D > E; E > F). In the latter variation, rats tested with (B vs. E) chose B more often than rats tested with (B vs. D). According to Van Elzakker et al. (2003) this finding is inconsistent with the logical inference account, which predicts no difference in (B vs.

D) and (B vs. E). Nevertheless, both the six-element and five-element results are consistent with their 'excitatory strength' theory, which argues that because E was trained against stimulus F (which was never reinforced) the animal simply learns to avoid F and thus assigns less excitatory value to E relative to D (see the value transfer theory below). Van Elzakker et al. (2003) suggest the possibility that the hippocampus is responsible for encoding the excitatory values rather than providing the neural substrate for the kind of relational comparison necessary for inference.

## The Value Transfer Theory

Allen (2006) provides a succinct description of the value transfer theory: "even though B and D are individually rewarded at the same rate, B is seen in association with A, which is always a winner. This is hypothesized to give B a positive boost in comparison to D" (p. 177). Whereas Van Elzakker et al. (2003) used a six-element paradigm to test this theory, Zentall (2001) describes an alternative approach that positively demonstrates this effect. Although Zentall used pigeons in his experiments rather than rats, the results are striking and deserve much greater attention than they have so far garnered. In the experiment two pairs of stimuli were differentially rewarded: stimulus A was rewarded 100% of the time over B that was never rewarded; and C was rewarded 50% of the time over D that was never rewarded (denoted $A_{100}B_0$; $C_{50}D_0$). When B was tested against D, B was preferentially selected, as predicted by the value transfer theory. Zentall concluded, "These results support value transfer theory and suggest that it may not be necessary to posit an ordered representation of the stimuli experienced during transitive-inference training" (p. 74).

Zentall (2001) describes further pigeon experiments that disclose the fact that the value transfer theory alone is inadequate and other quite subtle associative effects need to be accounted for in the context of the transitive inference paradigm. The value transfer theory predicts that just as A transfers positive value to B, B should also transfer *negative* value to A. To test for *negative* value transfer Zentall tested pigeons on ($A_{100}B_0$; $C_{100}D_{50}$; Test A vs. C). Negative value transfer predicts that C would be preferentially selected over A because the low-value B should 'drag down' the value of A by association, but Zentall found no such effect. Zentall then increased the subjects' experience with the negative stimuli B and D by reinforcing them in unpaired conditions: ($A_{100}B_0$; $C_{100}D_{50}$; $B_0$; $D_{50}$; Test A vs. C). However, instead of a preference of C over A as predicted by negative value transfer, the result was a preference for A over C! Zentall suggests that rather than a negative value transfer from B to A, the effect is the result of 'positive contrast.' That is, the value of A is enhanced by *contrast* with B (A = 100 vs. B = 0) as compared with the differential between C and D (C = 100 vs. D = 50). Testing ($A_{100}B_0$; $C_{50}D_0$; $B_0$; $D_0$; Test B vs. D) yielded preference for D over B indicating a *negative* contrast effect (i.e., B is much worse compared with A than D is compared with C).

The upshot of these results is that the five-element test paradigm is inadequate to yield conclusive evidence of inferential reasoning. The results so far are still too open to alternative plausible theoretical interpretations. Zentall's pigeon tests ought to be replicated in rats and should inspire further variations.

## Simulations

Relatively simple noncognitive systems can replicate the transitive inference effect, obviating the need to assume higher cognitive abilities such as rationality. De Lillo et al. (2001) showed that the five-element test results could be mimicked using a very simple back-propagating artificial neural network. Given this performance by a noncognitive system, De Lillo et al. suggest that "the binary, non-verbal, five-term-series task might not be suitable for detecting ontogenetic or phylogenetic trends in the development of the cognitive skills underlying inferential abilities" (p. 68). Furthermore, Frank, Rudy, and O'Reilly (2003) used artificial neural networks to model the hippocampus and their results supported Van Elzakker et al.'s (2003) excitatory strength theory. Overall, there exists reasonable doubt that the transitive inference effect as observed in rats is sufficient evidence of inferential reasoning, despite the observed role of the hippocampus in this effect.

## Causal Reasoning

The term 'causal reasoning' could be interpreted as a primitive ability to 'grasp' the causal power of objects in a nonconceptual sense (e.g., see Hoerl, 2011), such as that an event B is always followed by event A. This is a case of associative learning despite the temporal separation between the associated events and does not imply rationality. However, when the 'reasoning' in 'causal reasoning' is understood to involve inferential thinking, then rationality is implied. Certainly some researchers and commentators explicitly claim rationality in rats on this basis (e.g., Beckers, Miller, De Houwer, & Urushihara, 2006; Clayton & Dickinson, 2006; Mitchell, De Houwer, & Lovibond, 2009). In this section I analyze the results of causal reasoning experiments. I conclude that although reasonable cases can be made for 'causal reasoning' in the *nonconceptual* sense, this does not necessarily entail rationality in the terms I have delineated.

## Forward Blocking

Based on their experiments on Pavlovian conditioning in rats, Beckers et al. (2006) suggest that forward blocking is sensitive to constraints of causal inference. Forward blocking is a term describing the observation that once an unconditioned stimulus (US) has been paired with a conditioned stimulus (CS), in subsequent training that US will not be as strongly associated with a newly introduced and redundant CS that is presented alongside the original CS. In the nomenclature of comparative psychologists, call the first CS 'A'; the second 'X' and denote the US as '+'. The training is A+ followed by AX+. The observation is that X does not acquire much associative strength; in other words formation of the association between X and + is *blocked*. It is as if the subject infers that because A is already known to be the cause of the US, then X is probably not a cause of the US.

The forward blocking experiment is best explained starting with a human example. If Food A causes an allergic reaction and then Food A plus Food X causes the *same* reaction, then the inference is that since A is a known allergen X is probably *not* an allergen. Beckers et al. (2006) suggest that in humans forward blocking is the result of "effortful inferential reasoning"

(p. 93). To devise a rat analogue to the human case, Beckers et al. note that the assumed inference depends on certain constraints. The conclusion that X is not an allergen is reasonable if the subject assumes that (a) causes are *additive* such that two allergens in conjunction should cause a *greater* reaction than one alone; and (b) the ceiling (maximum possible reaction) in the AX+ regime has not been reached. In other words, there is a reason to believe that there ought to be a greater reaction if X is a contributor. Then, experiencing the effect of AX+ to be the same as A+ would suggest that X did not contribute to the effect. Given this, we can make testable predictions based on the difference between trials where the ceiling is not reached and those where it is. In the case where the ceiling is knowingly *not* reached subjects will infer that X is not a contributor (i.e., blocking will occur). In the case where the ceiling knowingly *is* reached subjects will not be able to definitively infer the causal status of X and therefore blocking should not occur. Beckers et al. (2006) tested these predictions in rat analogues of the human example given above and obtained positive results, tending to support a conclusion that rats are capable of causal inferential reasoning.

Experiment 1 in Beckers et al. (2006) aimed to train rats to learn that causes are nonadditive (see 'nonadditive' condition in Table 2). This was done in phase (i) by pairing individual cues (e.g., tones, buzzers and flashing lights) with a US (footshock) and also applying *combinations* of the cues with the same (not increased) level of shock (C+, D+, CD+). In phase (ii) the US was paired with a new stimulus (A+); in phase (iii) the US was paired with a combination of the phase (ii) stimulus and another new stimulus (AX+) (the control group was trained with B+ and then AX+). In the test phase the (thirsty) rats were tested against X alone with a lever present, which they had been pretrained to press for water rewards. The relative rate of lever pressing was used as a measure of the blocking effect: if the rats associated X with the US (i.e., blocking had *not* occurred) then they should display a fear effect in the test phase that would manifest as fewer lever presses. In effect, a relatively low rate of lever pressing indicates an inference that X was a cause of the footshock. As Table 2 shows, the results match predictions. In the nonadditive condition, both test and control groups press levers at a low rate: the rats

Table 2
*Beckers et al.'s (2006) Experiment 1*

| Condition | Group | (i) | (ii) | (iii) | Inference X = cause? | Expected effect: Blocking? | Relative lever pressing |
|---|---|---|---|---|---|---|---|
| Nonadditive | Test | C+ D+ CD+ | A+ | AX+ | ✓ | | ▮ |
| | Control | C+ D+ CD+ | B+ | AX+ | ✓ | | ▮ |
| Irrelevant element | Test | C+ D+ E+ | A+ | AX+ | | ✓ | ▮▮▮ |
| | Control | C+ D+ E+ | B+ | AX+ | ✓ | | ▮ |
| Irrelevant compound | Test | C+ C+ DE+ | A+ | AX+ | | ✓ | ▮▮▮▮ |
| | Control | C+ C+ DE+ | B+ | AX+ | ✓ | | ▮ |

have no reason to discount X as a possible cause. Contrast this with the other two conditions where the default assumption applies, that causes probably *are* additive. While in these cases the control groups cannot rule out X as a possible cause, the test groups *can* because they will 'reason' that as A is confirmed as a cause of the US (in phase ii), X cannot be a cause (else the footshock would have been of greater intensity).

In contrast to Beckers et al.'s (2006) first experiment, Experiment 2 was designed to train the rats to learn that causes *are* additive and Experiment 3 was designed to train the rats to learn that (in the test phase) the US ceiling had not been reached. Tables 3 and 4 show the results of Experiments 2 and 3, respectively. I will not analyze these here as similar principles

were used as for Experiment 1 and the data are self-explanatory. I leave it to the interested reader to verify that the results generally tend to support the hypothesis that rats are capable of inferring the causal status of X. The question I address now is whether the experimental results are sufficient to warrant the conclusion that rats are capable of inferential reasoning. Although the experimental design is ingenious and praiseworthy, I suggest that the results do not justify a conclusion of "symbolic causal reasoning processes" (Beckers et al., 2006, p. 100) in the rat subjects.

Because the results match the predictions, the key issue is whether the experiment itself is a valid indication of rationality. This experiment derives its legitimacy from replicating or approximating the actions of humans on the as-

Table 3
*Beckers et al.'s (2006) Experiment 2*

| Condition | Group | (i) | (ii) | (iii) | Inference X = cause? | Expected effect: Blocking? | Relative lever pressing |
|---|---|---|---|---|---|---|---|
| Additive | Test | C+ D+ CD++ | A+ | AX+ | | ✓ | ▮▮▮ |
| | Control | C+ D+ CD++ | B+ | AX+ | ✓ | | ▮ |
| Irrelevant element | Test | C+ D+ E++ | A+ | AX+ | | ✓ | ▮▮ |
| | Control | C+ D+ E++ | B+ | AX+ | ✓ | | ▮ |
| Irrelevant compound | Test | C+ C+ DE++ | A+ | AX+ | | ✓ | ▮▮▮ |
| | Control | C+ C+ DE++ | B+ | AX+ | ✓ | | ▮ |

Table 4
*Beckers et al.'s (2006) Experiment 3*

| Condition | Group | (i) | (ii) | (iii) | Inference $X =$ cause? | Expected effect: Blocking? | Relative lever pressing |
|---|---|---|---|---|---|---|---|
| Maximal | Test | ++ + | A++ | AX++ | ✓ | | |
| | Control | ++ + | B++ | AX++ | ✓ | | |
| Submaximal high | Test | ++++ ++ | A++ | AX++ | | ✓ | |
| | Control | ++++ ++ | B++ | AX++ | ✓ | | |
| Submaximal low | Test | ++ + | A+ | AX+ | | ✓ | |
| | Control | ++ + | B+ | AX+ | ✓ | | |

sumption that because humans' actions are inference-based then so too are the rats. We should exercise caution in this regard for at least two reasons. First, it is possible that the same results can occur due to one mechanism in humans (inference) and yet another mechanism (such as association) in rats. As Penn and Povinelli (2007) point out, the inapplicability of traditional associationist explanations does not rule out somewhat more sophisticated associative theories where there is no need to invoke inferential explanations (I discuss this further below). Second, it is possible that even the human actions were not attributable to inference despite the subjects' claims. Penn and Povinelli (2007) point out that inferential explanations of blocking effects in human studies rely on verbal self-reports that may be "post hoc redescriptions of effects initially generated through implicit nonpropositional mechanisms" (p. 102). In other words, humans may rationalize actions that were actually stimulated by nonpropositional mechanisms such as associations. For example, in the North et al. (1999) experiment mentioned earlier, the humans were almost all unaware of the influence of the background music on their choice of wine and gave spurious reasons for their choice. Thus, if the conclusion of inferential mechanisms in even the human version of the blocking experiment is in doubt the comparable results with rats should not be used to justify that same conclusion.

Another perhaps more compelling reason to question the validity of the Beckers et al. (2006) experiments is the fact that Haselgrove (2010) was able to reproduce the results using mathematical simulations based on the well-known Rescorla-Wagner associative model (Rescorla & Wagner, 1972). One potential problem with Haselgrove's simulation is that it assumes a common element in the cues used for pretraining and blocking, which was not explicit and (presumably) not intended in Beckers et al.'s set up. Haselgrove justifies this assumption on the basis that Beckers et al. did not demonstrate that the cues used were entirely different stimuli and point to the fact that five of the six used were of the same modality (auditory). I suggest that it is not unreasonable to assume that some constituents of the immediate environment could act as common elements, especially as the experiments were all conducted in identical operant chambers. In addition, even some researchers who question the validity of Haselgrove's assumptions (e.g., Guez & Stevenson, 2011) concede that the Beckers et al. (2006) results may well be within the explanatory power of associative models. For example, Schmajuk and Larrauri (2008) argue that causal learning is adequately explained by their attentional-associative model and they, too, provide evidence by way of mathematical models that replicate rat experimental results, implying that noncognitive mechanisms may be operative in rats. This possibility ought to be ruled out before we accept the strong conclusion of inferential thinking. Given this, the Beckers et al. (2006) results cannot be considered sufficient evidence of causal reasoning.

## Interventions

According to Blaisdell et al. (2006), in their experiments rats made causal *inferences* that cannot be explained by associative theories. Blaisdell (2009) suggests that these inferences are based on the *flexible use of representations* (p. 168), which seems to imply conceptual thought. Although Blaisdell stops short of *explicitly* claiming that rats are rational (A. Blaisdell, personal communication, October 18, 2011), it is clear that a case for rationality in rats can be made on the basis of his causal reasoning experiments, and there is no shortage of authors who explicitly do so (e.g., Clayton & Dickinson, 2006; Mitchell, De Houwer, and Lovibond, 2009).

Blaisdell et al. (2006) trained rats such that subsequent to a 10-s flickering light (L) they were presented with *either* a 10-s tone (T) *or* 10 seconds of sucrose delivery ('common cause'). The rats were also trained to associate a 10-s noise (*N*) with *simultaneous* delivery of 10 seconds of sucrose (F; see Figure 4). In the test phase of Experiment 1 rats were allocated to one of four conditions in which a lever was available, which was not available during the training. These conditions were intervene-T, observe-T, intervene-N, and observe-N. In the intervene-T condition, rats were presented with a 10-s T on pressing the lever, whereas in observe-T pressing the lever had no effect but T was presented 'randomly' (yoked to the Intervene-T lever). The same rules held for intervene-N and observe-N (see first three columns in Figure 5). The number of nose pokes into the food aperture was recorded as a measure of the rats' expectation of F.

According to Blaisdell et al. (2006), rats in the observe-T condition should reason that T was caused by L (but L was 'missed') so F should also be present, as F is also caused by L. By contrast, rats in the intervene-T condition should reason that because the cause of T was their own intervention (a lever press) rather than L, then F would not be present. If so, this predicts a lower rate of nose pokes for intervene-T than for observe-T. By contrast, there should be no difference in nose pokes between intervene-N and observe-N, as the rats would reason that N is a *direct* cause of F irrespective of lever presses. The average number of nose pokes per 10-s presentation of T or N is compared in the last column in Figure 5. The researchers found a significant difference in nose pokes between the intervene-T and observe-T conditions, but less so between the intervene-N and observe-N conditions, as predicted.

Blaisdell et al. claim that the results are inconsistent with associative theories but consistent with Bayes net predictions. Later, I discuss Bayes nets in the context of inferential reasoning, but first I consider how these results can be used to directly justify a claim of rationality. There are at least two areas in which explicit inferential reasoning can be invoked to explain the observed behavior. The first relates to the fact that in the test phase the rats searched for F when T occurred even though L was never presented, despite the fact that during training *either* T *or* F was presented following L but not both. Blaisdell (2009) explains the rats' expectation of F by reference to an acquired 'causal map' which he represents as: Tone ← Light → Food. Regarding the absence of L, Blaisdell et al. (2006) comment that "Apparently, in the initial phases of learning, rats tend to conservatively treat the absent but expected events as possibly present but missed" (p. 1021). Thus,
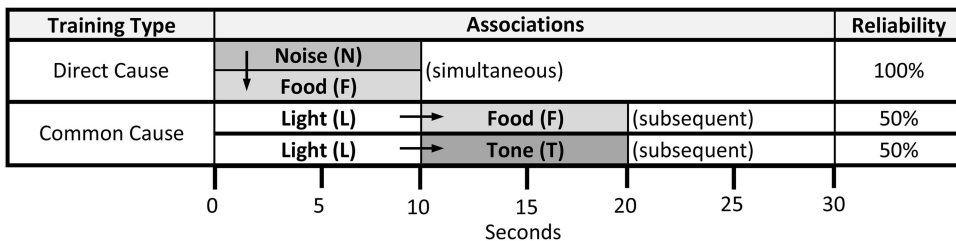
| Training Type | Associations | | | Reliability |
|---|---|---|---|---|
| Direct Cause | **Noise (N)** **Food (F)** | (simultaneous) | | 100% |
| Common Cause | **Light (L)** → **Food (F)** | (subsequent) | | 50% |
| | **Light (L)** → **Tone (T)** | (subsequent) | | 50% |

0    5    10    15    20    25    30
Seconds

*Figure 4.* Training phase of the causal reasoning experiments. *Note.* Rats were trained to expect food delivery (F) alongside a stimulus of noise (N); food delivery (F) subsequent to a flickering light (L); and tone (T) subsequent to flickering light (L).
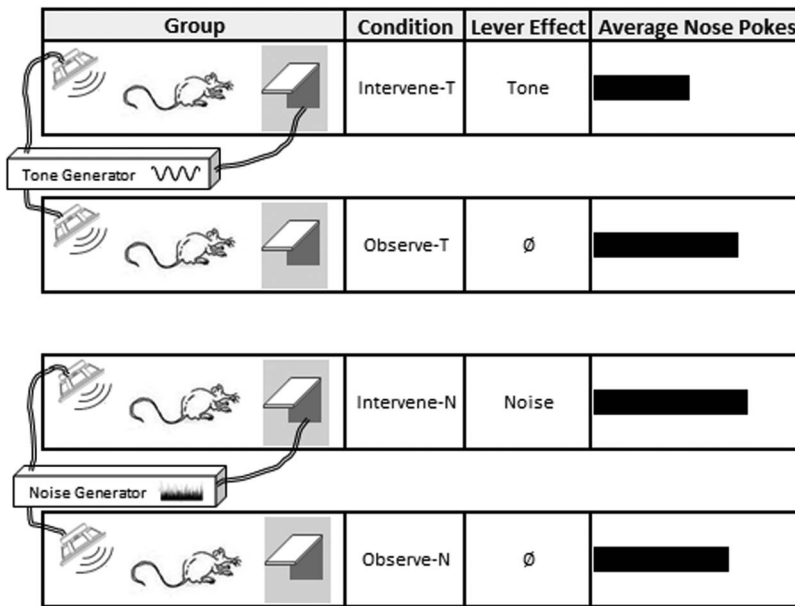
*Figure 5.* Test (extinction) phase of the causal reasoning experiment. Nose pokes in the Intervene-T condition are lower, as predicted.

the reasoning during the observe-T condition can be construed along the following lines:

Observe-T

1. L causes T
2. L (also) causes F
3. T occurred, therefore L must have occurred (and I must have missed it)
4. As L also causes F and L must have occurred, F will occur

The second case of possible inferential reasoning is in regard to the intervene-T condition in which the rat attributes the occurrence of T to its own lever-pressing action rather than to a 'missed' L. This can be construed along the following lines:

Intervene-T

5. T was caused by my intervention (lever press), not by L
6. As there was no L, F will not occur

I now raise several issues that cast substantial doubt on whether Blaisdell et al.'s (2006) experiments are sufficient evidence of rationality in rats. Several of these are similar to the objections raised to Beckers et al. (2006) above.

Blaisdell et al. compare rats' apparent ability to combine individually learned causal links into complex causal models with those abilities in humans, citing Perales, Catena, and Maldonado (2004). However, as argued above, even if (exclusively) rational processes were involved in the human case that does not guarantee that the same processes are involved in the rat case. Furthermore, as Perales et al. (2004) point out, data-driven (i.e., associative) and cognitively driven (i.e., causal-model based) theories can coexist as different learning strategies rather than alternative and mutually exclusive mechanisms. Again, if humans' results may be accounted for on the basis of associative mechanisms, the same may hold true for rats, rendering suspect the argument by analogy with humans as a sufficient basis for ascribing rationality to rats.

As can be seen in Figure 4, differences are apparent in the training regimes for different conditions. The direct cause association was trained by *simultaneous* presentation of the stimuli at 100% reliability. The common cause training associated the light/food and light/tone pairings at reliability of only 50% each and the presentations were not simultaneous but forward-paired.

It is difficult to predict how these differences might have affected the outcomes, but it might be significant enough to account for the failure to replicate the results in other labs (see below). As reviewed in the Transitive Inference section, there can be quite subtle and unexpected effects resulting from mediated conditioning and variable strength pairings. Indeed, Blaisdell himself agrees that repeating the experiment with properly aligned training regimes would be a worthwhile endeavor (A. Blaisdell, personal communication, October 19, 2011).

Plausible noncognitive accounts for the Blaisdell et al. results do exist, such as Delamater's (2011) 'limited capacity processing' and Dwyer, Starns, and Honey's (2009) 'response competition' account. Delamater suggests the possibility that lever press responses require processing within a limited capacity system in the rat. First, assume that the common cause training (see Figure 4) has led to an association between T and F via mediated conditioning. Then, it is possible that the processing required in the rat's brain for the lever press in the Intervene-T condition results in less effective processing of the subsequent T and this reduces the rat's ability to activate the association with F (and hence lower rate of nose pokes). The fact that this effect is not observed in the Intervene-N control groups could be accounted for if this limited capacity processing effect is greater for mediated conditioning compared with direct cause conditioning, a hypothesis that should be testable.

Delamater's (2011) suggestion is akin to Dwyer et al.'s (2009) theory that the reduced nose pokes in the Intervene-T condition is attributable to response competition between lever pressing and nose pokes. Dwyer et al. replicated the Blaisdell et al. (2006) experimental design[10] but although they also observed the intervene/observe difference, they note that this effect was at least as apparent in the direct cause condition as for the common cause condition; hence the findings were *not* replicated. Unlike Blaisdell et al. (2006), Dwyer et al. (2009) recorded the number of lever presses during presentation of stimuli. Their results show an inverse relationship between levels of lever pressing and magazine entries thereby supporting the view that the lower levels of lever pressing in the Intervene conditions were due to response competition rather than causal reasoning.

Even if Dwyer et al.'s (2009) idea of response competition is correct, this does not mean that rats are incapable of causal reasoning; perhaps in their experiments response competition masked the causal reasoning effect. To test this idea Burgess, Dwyer, and Honey (2012) designed several experiments to test for causal reasoning while minimizing response competition (e.g., by removing the lever immediately following a press or pretraining with a lever to avoid the effect of a novelty stimulus). Not only were Blaisdell et al.'s (2006) findings not replicated, neither were all of Dwyer et al.'s (2009). In particular, Burgess et al.'s (2012) Experiment 3 secured results that were the *exact opposite* of that predicted by causal map theory: Rats in the common-cause group responded *more* to T in the (equivalent of) the 'Intervene-T' condition than the 'Observe-T' condition.[11]

Differences in experimental set-up (e.g., rat species used, control of possible confounding factors) may account for these differing results, but this lack of robustness for the Blaisdell et al. (2006) results implies that those findings do not constitute reliable and therefore sufficient evidence of inferential reasoning in rats.

Mathematical models that replicate the Blaisdell et al. (2006) results cast further doubt on the reasoning ability hypothesis. Kutlu and Schmajuk (2012) successfully implemented a mathematical simulation of the causal reasoning experiments based on a 1996 model known as the attentional-associative model. Their model incorporates a mechanism that modulates attention to the CS in proportion to the total novelty detected in the environment and forms CS-CS and CS-US excitatory and inhibitory associations according to a real-time competitive rule. By adjusting the model's parameters—analogous to tweaking experimental factors such the choice of rat species—the mathematical simu-

---

[10] Not quite an exact replication: Dwyer et al. (2009) measured the amount of time spent in the magazine as compared with Blaisdell et al. (2006), who measured the number of entries into the magazine (nose pokes). Nevertheless, Dwyer et al.'s measure is appropriate because their hypothesis is that more time spent on one response (magazine entry) limits the amount of time available for the other response (lever pressing).

[11] I am using Blaisdell et al.'s nomenclature rather than Burgess et al.'s for consistency in the comparison.

lation of Blaisdell et al.'s experiments yields comparable results.

Blaisdell et al. suggested that an explanation should be sought compatible with Bayes net theories in lieu of associative theories, yet the two should not be considered mutually exclusive. Although at a high level Bayesian nets are intended to be models of (cognitive) causal inference (Pearl, 1997), at a neural level Bayesian processing might be how the brain implements learned associations. Bayes's theorem provides a mathematical procedure for calculating conditional probabilities, which begins with an assumed 'prior' probability of an event (say, the existence of food at a particular time and place) that is modified by subsequent incoming data to produce an updated 'posterior' probability. It is not unreasonable to assume that priors are determined in animals via their evolutionary history and individual previous experiences (McNamara, Green, & Olsson, 2006). Current experiences then provide the incoming data that can modify the prior to produce a posterior. In the context of rat experiments, the reinforcement schedule provides the incoming data. The fact that the rat's behavior changes as a function of learned associations is evidence of the derivation of a posterior. Conceivably, the posteriors are somehow encoded as combinations of action potentials at the neuronal level. Thus, associative and low level Bayesian-based explanations are not mutually exclusive.

From a 'top down' perspective, several other arguments give reason to doubt that Blaisdell et al.'s (2006) experiments are sufficient evidence of rationality in rats. If the inference in the Intervene-T condition was that 'T was not caused by L and therefore there would be no F,' then why were the nose-pokes not zero? Blaisdell suggests that "even if the rats thought it unlikely that food would be there they would have a look-see just to be sure" (A. Blaisdell, personal communication, October 18, 2011). Nevertheless, one must question how many such nose pokes would be conducted before a *rational* rat came to the conclusion that there was no point checking. To put the point another way, it is difficult to support a conclusion of rationality when the baseline is so arbitrary and the differences in above-baseline behavior so marginal. Of course we do expect the rat to go and check for food. As one commentator on

this paper remarked, the food giving apparatus is itself a stimulus. But that is an associative explanation, not one based on rationality. The same commentator remarked that the rats would be "stupid" (*sic*) to *not* check for food and I would not disagree—indeed, it is only if the rat were *rational* that it would *not* check for food. Or course, checking for food in this case does not therefore preclude the possibility that the rats are rational; they may be rational yet still go have a 'look-see.' My point is that the behavior does not constitute sufficient evidence of rationality. Note also that whether or not a rat *can* be "stupid" is precisely the question I am addressing because noncognitive systems can be neither smart nor stupid, only 'programmatic.' To call a creature stupid implies that it is capable of reasoning (faulty or not) in the first place.

Furthermore, note that the rats never grasp the fact that when L has caused T then F does not occur and vice versa. Blaisdell et al.'s (2006) explanation is that "With few learning trials, rats tend to integrate individual learning relations into a coherent integrated model. Only after many trials do rats encode the explicit absence of the nonpresented cues" (p. 1021). This seems to imply that to form an association between T and *lack of* F requires more extensive training. Whether true or not this is an associative type account and does not support a conclusion of rat rationality. Finally, in the observe-T condition why should we accept the supposition that the rats assume that L was present but simply missed? Are rational creatures likely to continue to make this error time and again?

Given the doubts raised above, these experiments cannot be considered sufficient evidence of rationality in rats. Rats do appear to perceive some 'causal' relations such as that *T follows L* and *F follows L* and *when the lever is pressed T occurs*, but there is no need to assume inferential capacities for this type of 'causal reasoning.' As mentioned earlier, Hoerl (2011) maintains that there are primitive abilities to 'grasp' the causal power of objects that do *not* imply concept possession. Indeed, alternative experimental paradigms aimed at investigating causal reasoning in rats supports the view that causal learning is explicable in an associative framework based on associations between contiguous

exogenous events (Polack, McConnell, & Miller, 2013).

## Goal Orientation

Since the early 1980s experiments on instrumental learning in rats have challenged the prevailing dogma that lever-press acquisition was controlled solely by sensorimotor learning involving a process of stimulus-response (S-R) association and instead suggested that animals are capable of a more elaborate form of encoding based on the response–outcome (R–O) association.[12] The S-R system is responsible for habitual behaviors whereas the R-O system is said to be responsible for goal-directed actions and these systems compete (and cooperate) in decision-making. Indeed, brain-imaging studies implicate differing regions in the human brain for these activities and homologous regions in rodent brains (Balleine & O'Doherty, 2010). Once again, given this correspondence between human and rat brains, and between some human and rat behavior, it is tempting to conclude that similar processing occurs in both species. To wit, because humans use prefrontal cortex in rational decision-making and rats employ a homologous brain region for apparently analogous behaviors, then it appears reasonable to claim that rats are similarly capable of rational thought (e.g., Dickinson, 1985).

Dickinson (1985) contends that at least some behavior in rats is under teleological control and cannot be explained at the psychological level in terms of internal associations:

> Rather, we argue that the knowledge about the action-goal relation must be encoded in a propositional-like form so that it can be operated on by a practical inference process to generate the instrumental performance. In this sense actions are inherently rational in a way that responses can never be. (Dickinson, 1985, p. 78)

Apparently Dickinson not only believes that rats are rational but that they entertain thoughts in propositional-like form. More recently, Dickinson (2012) has argued that a rat's 'practical reasoning' capacity emerges from associative processes though it is not clear how closely 'practical reasoning' matches our use of the term rationality. In fact, his description of cognition as arising from architectural constraints on associative processes seems to be a very mechanistic model. Nevertheless, Dickinson and Balleine's (2010) explanation for the results of goal-orientation experiments with rats assumes *conceptual* ability in the form of abstract representations of goal values (A. Dickinson, personal communication, January 31, 2013). I review these claims below.

### The Palermo Protocol

The best way to explain the significance of the rat experiment described below is to begin with the event that inspired it (Dickinson & Balleine, 2010). One day during a Sicilian holiday, Dickinson (AD) chanced upon a market selling watermelons and learned, first, how to subsequently navigate his way to the market, and second, that watermelon (never before tasted) is an excellent thirst-quencher. That evening he became horribly sick on the local red wine. A couple of days later, a thirsty AD again sought out the market but this time on sampling the watermelon he experienced nausea and the watermelon so disgusted him that he has not partaken since. AD later learned that a single pairing of a *novel* flavor with gastric sickness can condition an aversion to the flavor even after an hour or more. AD's sickness had been paired with the novel-flavored watermelon rather than the familiar red wine. What we can take from this story is that AD attained a conditioned aversion to watermelon *but he did not know it* until he went back for another taste. Thus it appears that a dual psychology is in operation: as a cognitive creature, AD's search for watermelon was a "*rational* and intentional action" (Dickinson & Balleine, 2010, p. 75, my emphasis) controlled by his belief and desire about watermelons. On the other hand, AD's nausea was a manifestation of nonrepresentational, reflex psychology.

---

[12] This view is still disputed. For example, Rescorla (1990) surmised that instrumental training may result in more elaborate hierarchical structures involving not just the stimulus (S) and the response (R) but also the outcome (O) in an analogous way to the standard S-R paradigm in a structure "of the form S-(R-O) in which the R-O association itself becomes associated with the stimulus" (p. 262). Rescorla successfully designed and tested S-(R-O) analogues to three standard Pavlovian phenomena (Kamin blocking effect; CS-US contingency; relative stimulus validity) and found results "consistent with the view that learning entails the development of an association between the stimulus and the R-O relation" (p. 269).

Balleine and Dickinson (1991) designed an experiment to test if rats, too, are subject to a dual psychology, which would imply that they are possibly capable of rationality. In the rat analogue the instrumental action is a lever press (in place of navigating to the market); the reinforcer is sucrose solution (in place of watermelon); and the latent aversion was induced by lithium chloride solution (instead of red wine). Each phase of the experiment was carried out on subsequent days:

(i) Rats were first given magazine training without levers present and then pretrained to lever press with a water reinforcer (reinforcers were presented throughout in the magazine); finally, lever pressing was extinguished.

(ii) All rats were trained in a single session to associate a sucrose solution with lever pressing; half the rats were then given an immediate injection of lithium chloride whereas the other half were given the injection after a delay of 6 hours.

(iii) Each of these groups was then further subdivided, half given access to sucrose solution and the other water, thereby creating four test groups in all.

(iv) In the test (extinction) phase, each group had access to a lever without reinforcement and lever pressing was counted.

(v) Finally, to establish that aversion had taken place, all rats were rewarded for lever-pressing with 30 presentations of water, after which the water presentation was replaced by sucrose.

The schedule is depicted on the left side of Table 5 (excluding the final reacquisition phase). The right side of Table 5 shows the results.

According to Dickinson (2008), the results of the experiment (and subsequent variations and extensions) replicate the events in the AD story. The IMM-H2O group acts like AD *before* he discovered the watermelon aversion: they lever-press at the same rate as the two delayed-injection groups (DEL-SUC and DEL-H2O) that experience no aversion. Just as AD sought out watermelon when thirsty, the IMM-H2O group act as if unaware of their latent aversion and hence press the lever in attempts to gain sucrose. This is the expected result, because that group has not been reexposed to sucrose following the lithium chloride injection. Conversely, the IMM-SUC group members *were* reexposed to sucrose and were therefore apparently consciously aware of their aversion to it. This group consequently ('rationally') chose to avoid sucrose (by avoiding lever pressing), like AD after he was reexposed to watermelon and consequently avoided future contact.

I begin my analysis with several pertinent observations. First, in the AD story the subject totally avoided the reinforcer following the reexposure, whereas in the experiment the relevant rats did not *totally* avoid pressing the lever but only (albeit, significantly) *reduced* lever-pressing. In the human case, if we interpret the *total* avoidance of the nausea-inducing substance as rational behavior, how should we interpret the contrasting rat case? The rat behavior does not *preclude* the existence of rationality, but the case for rationality is not as strong as it would be if the rats had totally avoided pressing the lever.

The second observation concerns the number of magazine entries following lever pressing (far right column in Table 5). The 'rationality' interpretation of the results relies on there being a difference between the IMM-SUC and IMM-DEL groups; however, the results for these groups are comparable if we compare magazine

Table 5
*Schedule for Experiment 1 in Balleine and Dickinson (1991)*

| | | (ii) | | | | Test results | |
| | (i) | | | (iii) | (iv) | Lever | Magazine |
| Group | Pre-training | Training | Aversion | Free access | Extinction | presses | entries |
|---|---|---|---|---|---|---|---|
| IMM-SUC | Lever → H2O | Lever → SUC | Immediate LiCl | SUCROSE | Lever → ∅ | 62 | 61 |
| IMM-H2O | Lever → H2O | Lever → SUC | Immediate LiCl | H2O | Lever → ∅ | 130 | 53 |
| DEL-SUC | Lever → H2O | Lever → SUC | Delayed-LiCl | SUCROSE | Lever → ∅ | 120 | 110 |
| DEL-H2O | Lever → H2O | Lever → SUC | Delayed-LiCl | H2O | Lever → ∅ | 132 | 148 |

entries rather than lever presses. The experimental analogue of AD's instrumental action (navigating to the market) was assumed to be lever pressing. But if magazine entry is considered to be the relevant analogue the results are consistent with an associative account. As discussed earlier in the Transitive Inference section, unpaired stimuli can be associated through mediated conditioning. Thus, the pairing (magazine → sucrose) coupled with the pairing (sucrose → sickness) should be sufficient to induce an association of the form (magazine → sucrose → sickness), which would then cause the rats to avoid entering the magazine to seek the expected sucrose.

Nevertheless, the above suggestion can only be part of the story because we require an additional account for the fact that the IMM-SUC group pressed the lever at a significantly lower rate. To account for the results of the Palermo experiment, Dickinson and Balleine (2010) argue that goal-oriented behavior is mediated by *goal value*, which according to Dickinson "ends up yielding an abstract representation of value (in the form of a desire) that can enter into inference processes, which is highly conceptual" (A. Dickinson, personal communication, January 31, 2013). Thus the argument for rationality hinges on the view that a rat's encoding of goal value is conceptual. I present an alternative account of 'goal value' based on the concept of 'affordance' that does *not rely on assuming conceptual abilities* in rats.

## Affordance

Affordance (Gibson, 1979) is a well-known theory developed to explain animals' perception of the 'values' of objects in their environment: "The *affordances* of the environment are what it *offers* the animal" (p. 127). There is no requirement for concept possession in affordance detection (e.g., "You do not have to classify and label things in order to perceive what they afford" (Gibson, 1979, p. 134). The idea of affordance has been debated among psychologists with some suggesting that there are different types, some of which actually do require conceptual abilities (Hartson, 2003; Norman, 1988). However, the notion of affordance that supposedly relies on conceptual abilities ('cognitive affordance') applies specifically to humans (and thus *presumes* the existence of con-

cepts). This notion of affordance primarily concerns the principles for appropriately designing devices for human usage. The type I lay claim to with regard to the case at hand—that is, *animal* detection of affordances—is called 'physical affordance' and does *not* rely on conceptual abilities. As earlier discussed, reasoning requires concept possession. Below I show that the rats' learning of goal values in the Palermo Protocol can be explained by the mechanism of physical affordance and so does not require the ability to reason.

Gibson emphasizes that in detecting an object's affordance the subject must learn its "true nature" (p. 142). For example: "consider substances that afford ingestion. Some afford nutrition for a given animal, some afford poisoning, and some are neutral" (p. 137). In the appropriate context, learning an object's true nature means learning its *goal value*. In the Palermo experiment the rat not only detects that a lever affords pressing but also learns the nature of what it delivers. All the rats learn the affordance of the lever to deliver an ingestible. Following re-exposure (phase iii) the IMM-SUC rats learn that the ingestible item (sucrose) affords poisoning and thus its negative goal value.

The above account based on affordance detection is at odds with Dickinson and Balleine's (2010) account only because of Dickinson's insistence that the rats encode 'desire'[13] in the form of an *abstract representation* of value. I agree that rats are most likely capable of mental representations,[14] but these do not necessarily need to be of an *abstract* form indicative of concept possession. 'Desires' that we can expect to find in a rat such as those elicited by thirst or hunger are visceral, not abstract. The notion of 'goal value' can be expressed in terms

---

[13] I use scare quotes for 'desire' here because in another context it is one of the propositional attitudes (including 'belief,' 'fear,' and so on) and this alone would presume propositional thoughts and hence concept possession. However 'desire' is commonly used in the context of animal behavior without assuming the existence of propositional thoughts.

[14] 'Mental representations' in this context need not imply abstract or intentional representations. For example, it is likely that rats entertain visual images of seen objects, which are examples of simple mental representations that need not involve concept possession (Savanah, 2012).

of 'strength of desire' with no need to assume that the rat possesses concepts.

I suspect that one major reason Dickinson is willing to attribute concept possession (and hence the potential for rationality) to rats is his belief that we only need to consider two broad levels: "S-R robots and cognitive creatures that are also endowed with intentional representations, affective experience, and the ability to integrate the two in consciousness" (Dickinson & Balleine, 2000, p. 201). The Model I propose includes an intermediate level (in which *affordance* is the operative mechanism) that is above the 'beast machine' though below the level of rationality. Compare this with Hurley (2003) who also sees behavioral flexibility as not an all-or-nothing capacity: "instrumental behavior is further along the spectrum in the direction of rational flexibility [than classical S-R behavior]" (p. 237). Hurley describes this level of agency as "flexible holistic relations between ends and means" (p. 237), which is strikingly reminiscent of how Gibson (1979, p. 143) describes affordance: the "inseparability" (holism) of the "possibilities of the environment" (the ends) and the "way of life of the animal" (the means). Later, Hurley (2006) talks of nonhuman animals as occupying 'islands of practical rationality' (note the qualifier, which distinguishes it from the notion of rationality I am discussing), which she describes as domain-specific reasons for action despite a *lack of conceptual abilities*.

## Future Research on Rat Rationality

To test for reasoning in animals, as discussed earlier, it will be necessary to eliminate behavior that can be accounted for as programmatic, such as species typical behaviors or behaviors acquired via associative learning. In addition, the test should be impervious to three prominent objections emerging from this study: (a) the inadequacy of arguments by analogy to human behavior or homologous human/rat brain regions; (b) the ability to replicate results using a noncognitive system such as a mathematical model; and (c) the existence of 'nonzero effects.' The experiment I propose to meet these criteria is the Staircase Test.

In essence, the Staircase Test requires rat subjects to fashion a tool to solve a novel problem that they would not encounter in their normal environments. The tool is simply two blocks of different size pushed together to create a climbing frame that allows the rat to reach a high platform containing a food reward. The idea is to train the subjects in the basic skills (pushing and climbing blocks of a certain size) and then test whether they can combine these skills in a novel (i.e., untrained) way to solve a problem (reaching food). The test meets our criteria because, in the first place, although the subject will apply actions that are already in its repertoire, the target behavior entails a complex *combination* of those actions that is neither species typical nor trained. Second, the test does not rely on comparison with analogous human behavior or brain regions. Third, I warrant that the behavior will not be replicable by a mathematical model (though of course this should be tested). Finally, to avoid the nonzero effects objection, I am content to set the base rate at zero and accept a single instance of success as sufficient evidence of rationality: A suitable configuration of the test set up will ensure a vanishingly low probability that the target behavior will be stumbled upon by random chaining of the trained actions (see *phase iv* description below).

I will rely on researchers to construct their version of the experiment according to best practices and pragmatic considerations. Accordingly, my explanation below includes only the key elements of the experiment and excludes considerations such as the reward presentation method, training session length and frequency, apparatus and chamber dimensions, and so forth.

### The Staircase Test

**Phase (i): Training.** Using a suitable reward to train the rats to push two different-sized blocks. Block 1 is more-or-less cube shaped, and Block 2 is a rectangular prism exactly double the height of Block 1 (see Figure 6, panel i). Block 1 should be of a size that just allows the rat to climb onto it (see next phase). Block 2 should stand like a column and designed to prevent toppling. Rats are rewarded for pushing either block by a minimum of a set distance.

**Phase (ii): Training.** Train the rats to climb on top of Block 1 *only* (see Figure 6, panel ii). The rats learn that they can reach up a certain height. Note that if Block 1 and Block 2
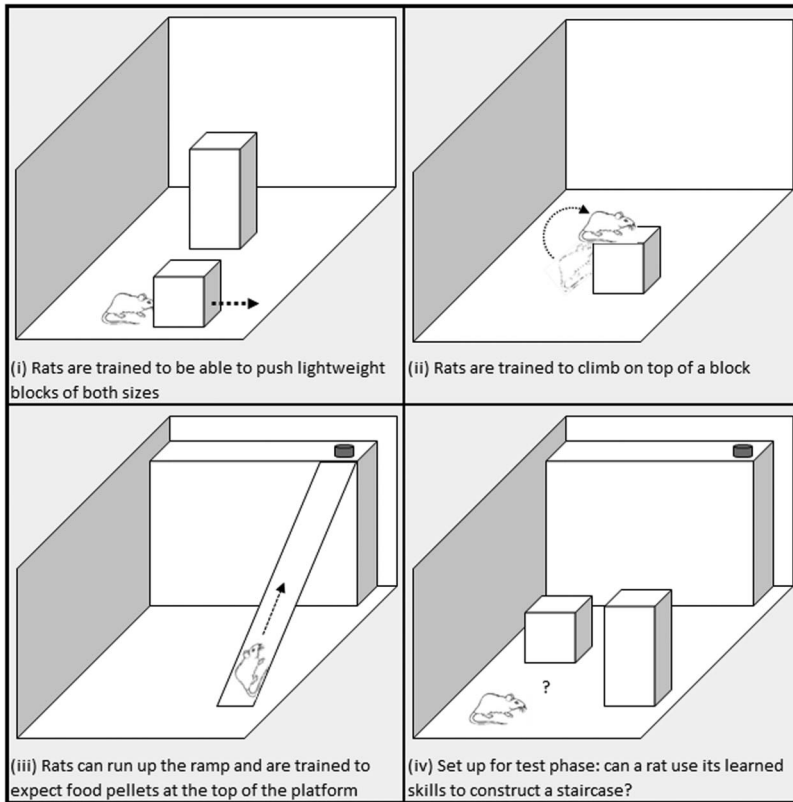
*Figure 6.* Schematic depiction of the Staircase Test. Panels (i) – (iii) are three separate training regimes. Panel (iv) shows the test phase set up.

are adjacent the height from Block 1 to the top of Block 2 is also climbable, though this is not specifically trained.

**Phase (iii): Training.** In a new chamber one wall has a high platform (exactly as high as three stacked Block 1s). The platform is accessible via a ramp (removable by the experimenter for the test phase). The rats are trained to expect food on the platform (see Figure 6, panel iii).

**Phase (iv): Test phase.** The trained rat is placed in the chamber, which contains Block 1 and Block 2 appropriately positioned to minimize chance success (i.e., distant from the each other and from the platform; see Figure 6, panel iv). A food reward is visible on the platform and the ramp has been removed. The question asked is *does the rat work out how to retrieve the food item?* (see Figure 7).

Despite the fact that this test sets the bar deliberately high, note that a skeptic (that is, an even greater skeptic than I) might *still* dispute that passing the Staircase Test is evidence of rationality. For example, it might be claimed that the rat detects the affordance of a climbing rig for reaching the platform. I would respond that this could only happen after *pushing the blocks together* and *positioning the whole staircase under the platform*, a planned combination of actions derivable only through inferential reasoning. Of course I would be more than pleased to engage in such a debate but I submit that the occasion will never arise as no rat will ever pass the test.

## Conclusion

Many researchers are willing to ascribe actual rationality to rats on the basis of the experiments analyzed herein. These researchers are to be commended on their many ingenious exper-
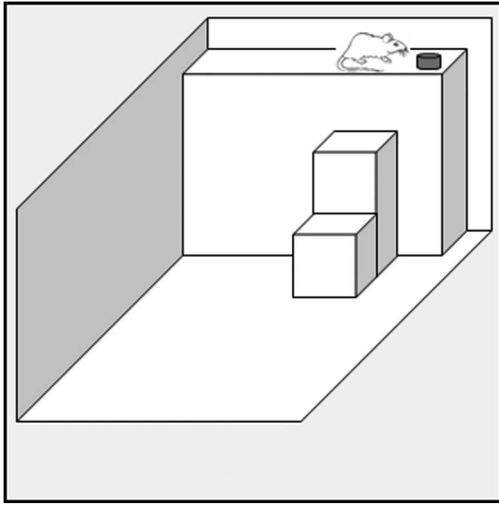
*Figure 7.* The goal of the Staircase Test. The rat has constructed a climbing rig to reach the food.

imental designs, but their results can be beguiling. Spatial navigation abilities are implemented probably by several different noncognitive mechanisms in different species. As such, rats' abilities in this regard remain inconclusive as demonstrations of rationality. The results of metacognition experiments using the bail-out paradigm are adequately explicable using traditional associative accounts. More complex associative theories are needed to account for some rat behavior. For example, Zentall (2001) demonstrated that there are many yet-to-be uncovered and subtle associative processes at work in the transitive inference effect.

A close examination of the causal reasoning experiments yields many issues that cast serious doubt on a conclusion that rats can reason. Probably the most convincing of all the paradigms explored here is Dickinson's Palermo Protocol (Dickinson & Balleine, 2010). Nevertheless, the level of cognition required to adequately account for the Palermo results need not extend to rationality; the now decades-old theory of affordance (Gibson, 1979) suffices.

Despite these misgivings, I expect that many researchers will remain convinced that their rats can reason. My challenge to them is the Staircase Test, the passing of which would convince me unequivocally of rat rationality.

## References

Allen, C. (2006). Transitive inference in animals: Reasoning or conditioned associations? In S. Hurley & M. Nudds (Eds.), *Rational animals?* (pp. 175–185). London, England: Oxford University Press.

Balleine, B., & Dickinson, A. (1991). Instrumental performance following reinforcer devaluation depends upon incentive learning. *The Quarterly Journal of Experimental Psychology Section B: Comparative and Physiological Psychology, 43,* 279–296.

Balleine, B. W., & O'Doherty, J. P. (2010). Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology, 35,* 48–69. http://dx.doi.org/10.1038/npp.2009.131

Beckers, T., Miller, R. R., De Houwer, J., & Urushihara, K. (2006). Reasoning rats: Forward blocking in Pavlovian animal conditioning is sensitive to constraints of causal inference. *Journal of Experimental Psychology: General, 135,* 92–102.

Blaisdell, A. P., Sawa, K., Leising, K. J., & Waldmann, M. R. (2006). Causal reasoning in rats. *Science, 311,* 1020–1022. http://dx.doi.org/10.1126/science.1121872

Blaisdell, A. (2009). The role of associative processes in spatial, temporal, and causal cognition. In S. Watanabe, A. P. Blaisdell, L. Huber, & A. Young (Eds.), *Rational animals, irrational humans* (pp. 153–172). Tokyo, Japan: Keio University Press.

Bunsey, M., & Eichenbaum, H. (1996). Conservation of hippocampal memory function in rats and humans. *Nature, 379,* 255–257. http://dx.doi.org/10.1038/379255a0

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,* 148–156. http://dx.doi.org/10.1037/a0027266

Campos, R., & Karmiloff-Smith, A. (2003). If metacognition exists in other species, how does it develop? Open Peer commentary on Smith et al. (2003), "The comparative psychology of uncertainty monitoring and metacognition." *Behavioral and Brain Sciences, 26,* 342. http://dx.doi.org/10.1017/S0140525X03240085

Carruthers, P. (1996). *Language, thought and consciousness.* Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511583360

Carruthers, P. (2003). Monitoring without metacognition. Open Peer commentary on Smith et al. (2003), "The comparative psychology of uncertainty monitoring and metacognition." *Behavioral and Brain Sciences, 26,* 342–343.

Carruthers, P. (2008). Meta-cognition in animals: A skeptical look. *Mind & Language, 23,* 58–89. http://dx.doi.org/10.1111/j.1468-0017.2007.00329.x

Cheng, K. (1986). A purely geometric module in the rat's spatial representation. *Cognition, 23,* 149–178. http://dx.doi.org/10.1016/0010-0277(86)90041-7

Clayton, N., & Dickinson, A. (2006). Rational rats. *Nature Neuroscience, 9,* 472–474. http://dx.doi.org/10.1038/nn0406-472

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8,* 1704–1711. http://dx.doi.org/10.1038/nn1560

Delamater, A. (2011). At the interface of learning and cognition: An associative learning perspective. *International Journal of Comparative Psychology, 24,* 389–411.

De Lillo, C., Floreano, D., & Antinucci, F. (2001). Transitive choices by a simple, fully connected, backpropagation neural network: Implications for the comparative study of transitive inference. *Animal Cognition, 4,* 61–68. http://dx.doi.org/10.1007/s100710100092

Dickinson, A. (1985). Actions and habits: The development of behavioral autonomy. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences, 308,* 67–78. http://dx.doi.org/10.1098/rstb.1985.0010

Dickinson, A. (2008). Why a rat is not a beast machine. In L. Weiskrantz & M. Davies (Eds.), *Frontiers of consciousness: Chichele lectures* (pp. 275–288). Oxford, UK: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199233151.003.0010

Dickinson, A. (2012). Associative learning and animal cognition. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences, 367,* 2733–2742.

Dickinson, A., & Balleine, B. W. (2000). Causal cognition and goal-directed action. In C. Heyes & L. Huber (Eds.), *The evolution of cognition* (pp. 185–204). Cambridge, MA: MIT Press.

Dickinson, A., & Balleine, B. (2010). Hedonics—The cognitive-motivational interface. In M. L. Kringelbach & K. C. Berridge (Eds.), *Pleasures of the brain*. Oxford, UK: Oxford University Press.

Dusek, J. A., & Eichenbaum, H. (1997). The hippocampus and memory for orderly stimulus relations. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 94,* 7109–7114. http://dx.doi.org/10.1073/pnas.94.13.7109

Dwyer, D. M., Starns, J., & Honey, R. C. (2009). "Causal reasoning" in rats: A reappraisal. *Journal of Experimental Psychology: Animal Behavior Processes, 35,* 578–586. http://dx.doi.org/10.1037/a0015007

Eichenbaum, H. (2000). A cortical-hippocampal system for declarative memory. *Nature Reviews Neuroscience, 1,* 41–50. http://dx.doi.org/10.1038/35036213

Fodor, J. (1975). *The language of thought*. New York, NY: Crowell.

Foote, A. L., & Crystal, J. D. (2007). Metacognition in the rat. *Current Biology, 17,* 551–555. http://dx.doi.org/10.1016/j.cub.2007.01.061

Frank, M. J., O'Reilly, R. C., & Curran, T. (2006). When memory fails, intuition reigns: Midazolam enhances implicit inference in humans. *Psychological Science, 17,* 700–707. http://dx.doi.org/10.1111/j.1467-9280.2006.01769.x

Frank, M. J., Rudy, J. W., & O'Reilly, R. C. (2003). Transitivity, flexibility, conjunctive representations, and the hippocampus. II. A computational analysis. *Hippocampus, 13,* 341–354. http://dx.doi.org/10.1002/hipo.10084

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.

Goel, V., Makale, M., & Grafman, J. (2004). The hippocampal system mediates logical reasoning about familiar spatial environments. *Journal of Cognitive Neuroscience, 16,* 654–664. http://dx.doi.org/10.1162/089892904323057362

Guez, D., & Stevenson, G. (2011). Is reasoning in rats really unreasonable? Revisiting recent associative accounts. *Frontiers in Psychology, 2,* 277. http://dx.doi.org/10.3389/fpsyg.2011.00277

Hampton, R. R. (2001). Rhesus monkeys know when they remember. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 98,* 5359–5362. http://dx.doi.org/10.1073/pnas.071600998

Hampton, R. R. (2005). Can rhesus monkeys discriminate between remembering and forgetting? In H. S. Terrace & J. Metcalfe (Eds.), *The missing link in cognition: Origins of self-reflective consciousness* (pp. 272–295). Oxford, UK: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195161564.003.0011

Hartson, R. (2003). Cognitive, physical, sensory, and functional affordances in interaction design. *Behaviour & Information Technology, 22,* 315–338. http://dx.doi.org/10.1080/01449290310001592587

Haselgrove, M. (2010). Reasoning rats or associative animals? A common-element analysis of the effects of additive and subadditive pretraining on blocking. *Journal of Experimental Psychology: Animal Behavior Processes, 36,* 296–306. http://dx.doi.org/10.1037/a0016603

Hoerl, C. (2011). Causal reasoning. *Philosophical studies. 152, 2,* 167–179.

Hurley, S. (2006). Making sense of animals. In S. Hurley & M. Nudds (Eds.), *Rational animals?* (pp. 139–

171). Oxford, UK: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780198528272.003.0006

Hurley, S. (2003). Animal action in the space of reasons. *Mind & Language, 18,* 231–256.

Jozefowiez, J., Staddon, J. E. R., & Cerutti, D. T. (2009). Metacognition in animals: How do we know that they know? *Comparative Cognition & Behavior Reviews, 4,* 29–39. http://dx.doi.org/10.3819/ccbr.2009.40003

Kacelnik, A. (2006). Meanings of rationality. In S. Hurley & M. Nudds (Eds.), *Rational animals?* (pp. 87–106). Oxford, UK: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780198528272.003.0002

King, J. (2003). Parsimonious explanations and wider evolutionary consequences. Open Peer commentary on Smith et al. (2003), "The comparative psychology of uncertainty monitoring and metacognition." *Behavioral and Brain Sciences, 26,* 347–348.

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,* 84–101. http://dx.doi.org/10.1037/a0026221

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,* 686–708. http://dx.doi.org/10.1037/a0026478

McNamara, J., Green, F., & Olsson, O. (2006). Bayes' theorem and its applications in animal behavior. *Oikos, 112,* 243–251. http://dx.doi.org/10.1111/j.0030-1299.2006.14228.x

Metcalfe, J. (2003). Drawing the line on metacognition. Open Peer commentary on Smith et al. (2003), "The comparative psychology of uncertainty monitoring and metacognition." *Behavioral and Brain Sciences, 26,* 350–351.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences, 32,* 183–198. http://dx.doi.org/10.1017/S0140525X09000855

Mitchell, R. W. (2002). Subjectivity and self-recognition in animals. In M. R. Leary & J. Tangney (Eds.), *Handbook of self and identity* (pp. 567–593). New York, NY: Guilford Press.

Moser, E. I., Kropff, E., & Moser, M.-B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annual Review of Neuroscience, 31,* 69–89. http://dx.doi.org/10.1146/annurev.neuro.31.061307.090723

Nadel, L., & MacDonald, L. (1980). Hippocampus: Cognitive map or working memory? *Behavioral & Neural Biology, 29,* 405–409. http://dx.doi.org/10.1016/S0163-1047(80)90430-6

Norman, D. (1988). *The psychology of everyday things.* New York, NY: Basic Books.

North, A., Hargreaves, D., & McKendrick, J. (1999). The influence of in-store music on wine selections. *Journal of Applied Psychology, 84,* 271–276. http://dx.doi.org/10.1037/0021-9010.84.2.271

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as cognitive map.* Oxford, UK: Clarendon Press.

Pearl, J. (1997). Bayesian networks. In R. Wilson & F. Keil (Eds.), *The MIT encyclopedia of the cognitive science* (pp. 72–74). Cambridge, MA: MIT Press.

Penn, D. C., & Povinelli, D. J. (2007). Causal cognition in human and nonhuman animals: A comparative, critical review. *Annual Review of Psychology, 58,* 97–118. http://dx.doi.org/10.1146/annurev.psych.58.110405.085555

Perales, J., Catena, A., & Maldonado, A. (2004). Inferring non-observed correlations from causal scenarios: The role of causal knowledge. *Learning and Motivation, 35,* 115–135. http://dx.doi.org/10.1016/S0023-9690(03)00042-0

Polack, C., McConnell, B., & Miller, R. R. (2013). Associative foundation of causal learning in rats. *Learning & Behavior, 41,* 25–41. http://dx.doi.org/10.3758/s13420-012-0075-5

Rescorla, R. A. (1990). The role of information about the response-outcome relation in instrumental discrimination learning. *Journal of Experimental Psychology: Animal Behavior Processes, 16,* 262–270. http://dx.doi.org/10.1037/0097-7403.16.3.262

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. K. Prokasy (Eds.), *Classical conditioning* (Vol. 2, pp. 64–99). New York, NY: Appelton-Century-Crofts.

Rödl, S. (2007). *Self-consciousness.* Cambridge, MA: Harvard University Press.

Savanah, S. (2012). The concept possession hypothesis of self-consciousness. *Consciousness and Cognition, 21,* 713–720. http://dx.doi.org/10.1016/j.concog.2011.02.019

Savastano, H. I., & Miller, R. R. (1998). Time as content in Pavlovian conditioning. *Behavioural Processes, 44,* 147–162. http://dx.doi.org/10.1016/S0376-6357(98)00046-1

Schmajuk, N., & Larrauri, J. (2008). Associative models can describe both causal learning and conditioning. *Behavioural Processes, 77,* 443–445. http://dx.doi.org/10.1016/j.beproc.2007.09.010

Schwitzgebel, E. (2011). *Belief.* In E. N. Zalta (Ed.), The Stanford encyclopedia of philosophy (Winter 2011 ed.). Retrieved from http://plato.stanford.edu/archives/win2011/entries/belief/

Shettleworth, S. J., & Sutton, J. E. (2003). Animal metacognition? It's all in the methods. Open Peer commentary on Smith et al. (2003), "The comparative psychology of uncertainty monitoring and metacognition." *Behavioral and Brain Sciences, 26,* 353–354.

Smith, J. D., Shields, W. E., & Washburn, D. A. (2003). The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences, 26,* 317–339. http://dx.doi.org/10.1017/S0140525X03000086

Smith, J. D. (2005). Studies of uncertainty monitoring and metacognition in animals and humans. In H. S. Terrace & J. Metcalfe (Eds.), *The missing link in cognition: Origins of self-reflective consciousness* (pp. 242–271). Oxford, UK: Oxford University Press.

Tarsitano, M. S., & Jackson, R. R. (1997). Araneophagic jumping spiders discriminate between detour routes that do and do not lead to prey. *Animal Behaviour, 53,* 257–266. http://dx.doi.org/10.1006/anbe.1996.0372

Van Elzakker, M., O'Reilly, R. C., & Rudy, J. W. (2003). Transitivity, flexibility, conjunctive representations, and the hippocampus. I. An empirical analysis. *Hippocampus, 13,* 334–340. http://dx.doi.org/10.1002/hipo.10083

Zentall, T. R. (2001). The case for a cognitive approach to animal learning and behavior. *Behavioural Processes, 54,* 65–78. http://dx.doi.org/10.1016/S0376-6357(01)00150-4

Zentall, T. (2003). Evidence both for and against metacognition is insufficient. Open Peer commentary on Smith et al. (2003), "The comparative psychology of uncertainty monitoring and metacognition." *Behavioral and Brain Sciences, 26,* 357–358.