On Becoming Approximately Rational: The Relational Reinterpretation hypothesis

by Derek C. Penn and Daniel J. Povinelli

Derek C. Penn, dcpenn@ucla.edu (corresponding author)
Daniel J. Povinelli, ceg@louisiana.edu
Cognitive Evolution Group
University of Louisiana at Lafayette
New Iberia, LA, 70560 USA
http://www.cognitiveevolutiongroup.org/

Abstract

Some of the most contentious and intractable debates in comparative psychology result from the fact that researchers persistently overlook—or frankly refuse to acknowledge—the distinct roles played by different levels of explanation in cognitive science. We examine two prominent explananda—"Do animals reason about causal relations?" and "Do animals have a theory of mind?"—and show how in each case, comparative researchers continue to gloss over the difference between functional- and representational-level claims. We propose a representational-level hypothesis about how nonhuman animals have evolved to become *approximately* rational without employing higher-order reasoning processes. And we show how comparative researchers might become approximately rational as well.

1 Introduction

The fact that an organism behaves in a manner consistent with some "rational" or "folk psychological" model of cognition does not mean that the organism necessarily represents or reasons about the entities, variables and relations posited by that model. This point would be incredibly boring were it not completely ignored—even dismissed— by most comparative cognitive psychologists. Indeed, as we will show in this chapter, some of the most contentious and intractable debates in comparative cognitive research result from the fact that comparative researchers persistently overlook—or frankly refuse

to acknowledge—the distinct roles played by different levels of explanation in cognitive science.

We are hardly the first to make this point (see, for example, Bermudez, 2003; Dennett, 1987; Kacelnik, 2006; Shettleworth, 1998). But perhaps we can make the point so forcefully this time that nobody else will ever have to make it again. One can always be hopeful.

Our plan of attack is as follows: First, we review the basics: What is a "functional" model of cognition? How does this differ from a representational-level explanation? What is the relationship between the two? Second, we take two prominent and contentious explananda in comparative research—"Do animals reason about causal relations?" and "Do animals have a theory of mind?"—and show how in each case, comparative researchers continue to gloss over the distinct roles that functional- and representational-levels of explanation play in answering these explananda. Finally, we propose a representational-level hypothesis about how nonhuman animals have evolved to *approximate* a minimal rational model of causal reasoning and theory of mind. And we show how comparative researchers might become approximately rational as well.

2   The Etiology of a Muddle

Let us define a *functional-level* explanation as any explanation that specifies how a cognizer will or should behave given a certain set of inputs—and provides some insight into *why* the cognizer should behave in this way—without making any claims or assumptions about how that cognizer is representing or carrying out the given computations or inferences. And let us define a *representational-level* explanation as any explanation that specifies the representations and relations that play a causally efficacious role in a cognizer's computations and makes some claim about how those computations are actually carried out in the cognizer's head[1].

Our definition of a functional-level explanation encompasses both a "computational-level" explanation in Marr's (1982) sense and a "rational model" in Anderson's (1990) sense, as well as Dennett's (1987) "Intentional" and "Design" stances and Kacelnik's (2006) "economic" and "biological" concepts of rationality. Our

definition of a representational-level explanation encompasses both Marr's "algorithmic" and "representational" levels, Kacelnik's "psychological" level of rationality, and most of the intractable debate between symbolic and connectionist models of cognition (e.g., Fodor & Pylyshyn, 1988; Shanks, 2005).

As Dennett (1987) sagely pointed out, there is rarely, if ever, an isomorphic relationship between the entities, variables and relations posited by a functional model of a cognitive behavior, and the representations and relations that actually play a causally efficacious role in the computations carried out by the organism in question (see also Anderson, 1990; Marr, 1982). Nearly everyone agrees, at least in principle, that functional-level claims should be made on an implicit basis: i.e., the cognizer is postulated to behave *as if* it were making the computations specified by the functional-level model. And nearly everyone agrees, at least in principle, that representational-level models should attempt to explain how a system *approximates* the behavioral specifications laid out by a functional-level model without expecting there to any straightforward mapping between the two models.

But what comparative psychologists agree to in principle and what they do in practice are often quite different. Many of the most intractable debates in comparative psychology result from the fact that researchers persistently misinterpret functional-level models as representational-level claims. You can tell when a researcher has slipped into a representational-level claim when she argues or implies that the entities, variables and relations posited in her functional-level model actually play a causally efficacious role in the computations and/or psychological dynamics of the subject's cognitive system.

Kacelnik (2006) nicely straightens out the mess in the case of foraging among birds. We'll try to do the same for two other prominent comparative explananda.

3   Do Rats Reason About Causal Relations?

We have previously hypothesized that nonhuman animals are able to represent and reason about first-order causal relations—including the effect of their own instrumental actions—but that they lack the ability to represent or reason about higher-order causal relations and unobservable physical mechanisms (see Penn et al., 2008a;

Penn & Povinelli, 2007a; Povinelli, 2000). In the present section, we reexamine our hypothesis in light of a set of seminal experiments conducted by Blaisdell and colleagues (Blaisdell et al., in press; Blaisdell et al., 2006; Leising et al., 2008). Blaisdell and colleagues have interpreted their results as challenging our representational-level hypothesis about nonhuman causal cognition (see Leising et al., 2008; Waldmann et al., 2008). Hereinbelow, we show how a minimal rational model of Blaisdell et al.'s (2006) results is, in fact, consistent with our hypothesis and inconsistent with Blaisdell et al.'s original claims.

3.1   Rats and Causal Models

Blaisdell et al. (2006) presented rats with pairs of stimuli purportedly corresponding to one of two alternative causal structures. Rats presented with a *common-cause* model were given pairings of a light, *L*, followed by a tone, *T*, and, separately, the same light, *L*, followed by a food reward, *F*. These rats were also trained on a *direct-cause* model in which a noise, *N*, was presented simultaneously with food. Rats presented with a *causal-chain* model were given pairings of *T* followed by *L* and then, *L* followed by *F*.

[Insert Figure 1 Here]

During the test phase, the rats in the common-cause condition were divided into one of four test conditions. Rats in the two *Intervene* conditions received a presentation of *T* or *N*, respectively, each time they pressed a lever in the test chamber. Rats in the two *Observe* conditions observed presentations of *T* or *N* that were independent of their own actions on the lever. The experimenters recorded the number of nose pokes the rats made into the magazine where *F* had been delivered during the training phase.

Blaisdell et al. (2006) found that rats in the *Intervene* condition who had been trained on the common-cause model made fewer nose pokes than rats in *Observe* condition. In contrast, there was no significant difference between the *Intervene* and *Observe* conditions between subjects in the causal-chain or within-subject for the direct-cause group.

Blaisdell et al. (2006) originally interpreted their results as consistent with a causal Bayes net account of causal reasoning in rats. Causal Bayes net theories posit that subjects reason *as if* they are forming integrated, structured graphical models of the conditional dependencies among causes and effects (e.g., Gopnik et al., 2004). One of the key assumptions of a causal Bayes net account is that cognizers should act *as if* they are sensitive to the causal Markov condition. The causal Markov condition says that if one holds all the direct causes of a given variable constant, then that variable will be statistically independent of all other variables in the causal graph that are not its effects.

Causal Bayes net theories are clearly *functional*-level explanations of causal reasoning. For example, as Steyvers et al. (2003) explain:

> Our models attempt to explain people's behavior in terms of approximations to rational statistical inference, but this account does not require that people actually carry out these computations in their conscious thinking, or even in some unconscious but explicit format. A plausible alternative is that people follow simple heuristic strategies, which effectively compute similar outputs as our rational models without the need for any sophisticated statistical calculations.

Blaisdell et al. (2006), however, slip into a representational-level interpretation of causal model theory without seeming to notice it. In their original paper, for example, Blaisdell et al. (2006) interpret their results as showing that "the core competency of reasoning with causal models seems to be already in place in animals" (p. 1022). Later, Leising et al. (2008) argue that the object of these experiments was to test whether rats "have acquired representations of causal models rather than merely associative knowledge" (p. 514) and claim that rats have "integrated" individual associations (i.e., light-tone, light-food) into a single "common-cause model" (p. 515).

In part because they have conflated functional- and representational-level claims, Leising et al. (2008) assume that their results "contradict" (p. 524) the account of nonhuman causal reasoning presented by Penn and Povinelli (2007a). In Penn and Povinelli (2007a), we argued that both human and nonhuman animals are "sensitive to the unobservable constraints specific to causal inference" (p. 102) and are able "to derive novel interventional predictions from purely observational learning" (p. 106). But we

contended that Blaisdell et al.'s (2006) results provide no evidence that rats are "sensitive to the causal Markov condition" (p. 106) or are able "to cognize their own interventions in an epistemic fashion" (Penn & Povinelli, 2007a, p.106). And we hypothesized that nonhuman animals are incapable of forming the kind of integrated, systematic, higher-order representations necessary to reason with causal models in their head (see also Penn et al., 2008a). Leising et al. (2008) argue that their results weaken our claim that "there is a sharp dividing line with respect to causal reasoning between human and nonhuman animals" (p. 526).

Ironically, the most compelling criticism of Blaisdell et al.'s (2006) and Leising et al. 's (2008) claims now comes from some of the original authors themselves (i.e., Waldmann and Blaisdell).

3.2   The Single Effect Learning Model as a Minimal Rational Model

Waldmann et al. (2008) argue that cognitive scientists should "consider whether there are alternative rational theories which are less computationally demanding while still fully accounting for the data." And Waldmann et al. propose a methodological heuristic they call the "minimality" requirement: *Ceteris paribus*, a rational model that is 1) more consistent with the relevant psychological evidence and the computations an organism can actually accomplish and 2) requires less computational complexity is to be preferred over a equally rational model that is less consistent with the psychological evidence or computational capacities of the organism or requires greater computational complexity.

Waldmann et al. (2008) propose a minimal rational explanation of the rats' behavior in Blaisdell et al. (2006) based on Buehner and Cheng (2005)'s "single-effect learning" model. According to this model, the simple common-effect structure is the basic unit in which causal learning occurs and organisms focus primarily on evaluating single causal relations during learning (see also Cheng, 1997). When evaluating the causal power of any given candidate cause, c, the learner groups all other potential causes of the given effect into a single variable, a, and assumes, as a simplifying heuristic, that a and c operate on their common effect independently. When confronting problems which involve traversing multiple causal links, the single-effect rational learner does not form

integrated representations of a causal model but, instead, makes inferences consecutively across causal relations by working from one link to the next via their common variables.

According to the single-effect learning model, Blaisdell's rats focused on single causal relations (e.g., light-tone or light-food) during the learning phase of this experiment without updating or representing an integrated common-cause model. When rats in the *Intervene* condition of the common-cause group produced a tone after pressing on a lever, the act of intervening on the lever deterministically implies that the tone need not be explained by other causes. This led the rats to act *as if* they were discounting the occurrence of the previously observed cause (light) and hence to lower their expectation of food. Conversely, upon hearing the tone, rats in the *Observe* condition of the common-cause group acted *as if* they were "diagnostically" inferring light even though they did not actually Observe light. Postulating light, the rats then proceed in a predictive direction to infer food from light.

Waldmann et al. (2008) argue that the single-effect learning model is more consistent with the rats' behavior than the causal Bayes net account originally advocated by Blaisdell et al. (2006). We believe they are right. As we pointed out previously (Penn & Povinelli, 2007a), a causal Bayes net model has great difficulty explaining how the rats inferred a common-cause or a causal-chain structure in this experiment given the fact that the data the rats actually observed violated the causal Markov condition (i.e., T and F were not independent conditional on the state of L). Because the single-effect learning model postulates that rats learn about separate links without representing how indirectly linked elements are related to each other, the single effect learning model does not require the causal Markov condition and accounts for both the learning and test phases of this experiment quite nicely.

3.3   Evaluating the Representational-Level Complexity of Rational Models

Waldmann et al. acknowledge that "a rational model for the aplysia will surely look different from one for humans." But they make no mention of how to formally evaluate the biological or psychological plausibility of competing rational models for different kinds of cognizers. In particular, though they mention, in passing, that biological brains may have different computational capabilities and constraints than digital

computers, there is no mention of the particular kinds of constraints that biological brains might have relative to digital computers. Ironically, because they do not take into account any representational-level concerns, Waldmann et al. (2008) do not provide a formal argument for why their single-effect learning model is less computationally demanding or more biologically plausible than a causal Bayes net alternative. Fortunately, we believe there is a good one.

There are good, principled reasons for believing that structured relations are more difficult to implement in a biological brain than unstructured associations, and that higher-order structured relations are more computationally demanding than lower-order structured relations (Halford et al., 1998; Holyoak & Hummel, 2000). Thus, a metric such as "relational complexity" provides a useful measure of the complexity of competing rational models for biological cognizers. The relational complexity of a problem is a function of the number of structured relations that must be integrated simultaneously by the cognizer (see Halford et al., 1998 for details). Relational complexity has already been used quite successfully to explain psychological, neural and developmental effects in ToM, causal learning and many other forms of relational reasoning (e.g., G. Andrews & Halford, 2002; Badre, 2008; Bunge et al., 2005; Christoff & Keramatian, 2007; Halford et al., 2002; Kroger et al., 2004). It also accords well with what we know about the challenges of approximating relational inferences in a neurally plausible computational architecture (Doumas & Hummel, 2005; Holyoak & Hummel, 2000; Hummel & Holyoak, 2005). Importantly, relational complexity is quite distinct—and often orthogonal—to other more general measures of cognitive complexity, such as informational complexity in Shannon and Weaver's (1949) sense.

The relational complexity of a problem is also not reducible to the number of nodes or links involved. A causal problem with hundreds of relations but in which each individual relation can be considered sequentially has a lower effective relational complexity than one in which multiple relations must be integrated simultaneously.

We thus propose the following extension to Waldmann et al.'s (2008) minimality heuristic: *Ceteris paribus,* a rational model that requires less relational complexity is to be preferred over one that requires more.

A relational complexity metric provides strong support for the single-effect learning model as a minimal rational model of nonhuman causal reasoning. In a single-effect learning model, cognizers make inferences by evaluating one relation at a time and by consolidating information about multiple independent causes into a single variable. This is exactly the strategy that Halford et al. (1998) suggested biological cognizers employ for reducing the effective complexity of relational problems: i.e., divide up the problem into pieces that can be solved *sequentially* and then *chunk* all background relations not currently being considered into a single variable. The single-effect learning model shows how a cognizer can make rational causal inferences without needing to integrate multiple causal relations or forming higher-order relational representations.

Moreover, there are strong empirical reasons for believing that nonhuman animals are incapable of integrating multiple relations or representing higher-order relations (see Penn et al., 2008a). Thus, Buehner and Cheng's (2005) single-effect learning model is not only more consistent with the behavioral data than a causal Bayes net account, it also has lower effective relational complexity and is more consistent with the broader empirical evidence concerning nonhuman animals' cognitive capacities.

3.4   But Rats Are Only Approximately Rational...

The fact that the single-effect learning model is a compelling minimal rational model for the rats' behavior in Blaisdell et al. (2006) does not imply, of course, that this is a model of the representational-level processes employed by these rats. Indeed, the evidence suggests that the representational-level mechanisms rats are actually using sometimes cause them to deviate, quite significantly, from the predictions of this rational model.

For example, rats in the common-cause *Observe* condition expected food upon hearing the tone even though light—the common cause—was absent. In order to make the rats' behavior consistent with a rational model of causal cognition, one must postulate some additional psychological anomaly or limitation. The explanation favored by both Blaisdell et al. (2006) and Waldmann et al. (2008) is that the rats simply didn't notice the absence of light and/or suffered from some kind of memory deficit.

This hypothesis is difficult to sustain. The light stimulus was produced by turning off the house light and flashing a diffuse light (2/s) for 10 seconds. The rats were quite attentive to the presence or absence of this flashing light stimulus in other parts of the same experiment and had no problem remembering quite complex patterns of stimuli. Indeed, when rats in the causal-chain group were trained on a tone followed by light and then, separately, on a light followed by food, they did not show second-order conditioning unless the light bulb was hidden behind a cover (see also Blaisdell et al., in press). It seems worth considering whether there might be some other reason why rats in the common-cause *Observe* condition inferred food given tone even in the glaring absence of the common-cause, light.

3.5   How To Build a Rat That Approximates a Single-Effect Learning Model
Let us assume that the cognitive architectures of all animals (even slugs and humans) are capable of associative learning but that many animals (e.g., at least vertebrates) have evolved the *additional* ability to evaluate the causal power of an association using the constraints specific to causal relations and to represent causal relations in a functionally compositional fashion (see Cheng, 1997; and our discussion in Penn & Povinelli, 2007a). What further representational abilities do we need to add in order to explain the behavior of the rats in Blaisdell et al. (2006)?

Not many. We simply need to postulate that rats' system for causal reasoning has been designed so that the greater the causal power of a relation, the more an intervention on a causal node elicits an expectation of the corresponding effect and, conversely, that the lower the causal power of a relation, the more interventions interfere with an expectation of the effect. We also need to assume that rats do *not* possess the representational architecture necessary to represent and reason about higher-order, role-based relations, to integrate multiple relations, or to access relations in a systematic and omnidirectional fashion.

All of the above assumptions are consistent with our "Relational Reinterpretation" (RR) hypothesis (see Penn et al., 2008a, for details). We will now argue that this representational-level hypothesis explains *how* the rats in Blaisdell et al. (2006)

approximated a single-effect learning model and also *why* they sometimes deviated from this rational model.

Exposed to pairings of light-tone and light-food, our RR hypothesis postulates that rats in the common-cause condition formed two distinct representations of the causal relation between the paired stimuli. The two relations are linked via their common element but there is no integrated, relational representation of a higher-order "common-cause" structure. The rats also formed an indirect association via second-order conditioning between tone and food as predicted by Yin et al. (1994). Unlike the associations between light-tone and light-food, however, this indirect association lacked causal power (Given many more trials, the rats would have presumably formed an indirect, inhibitory association between the two stimuli since tone and food were negatively correlated with each other). In the causal-chain condition, rats also formed two distinct representations of the causal relations linked together via the common middle term and formed an indirect association between light and food via second-order conditioning as well.

According to our RR hypothesis, rats in the *Observe* common-cause group expected food after observing tone as a result of the second-order excitatory association. In the case of the *Intervene* common-cause group, rats did not expect food after pressing the lever because the link between tone and food was non-causal and, ex hypothesi, producing a tone by intervention suppressed (or competed with) the non-causal association with food. In the case of the *Observe* and *Intervene* causal-chain groups, observing or producing a tone both elicit a causal expectation of light and an expectation of light elicits a causal expectation of food because the two relations in this case had significant causal power. Ex hypothesi, an intervention on the causal node of a relation with significant causal power does *not* interfere with an expectation of the effect.

Representational-level hypotheses are only interesting if they show how a cognizer's behavior will deviate from the rational model it is approximating. Ours does. Contrary to a rational model of causal reasoning, the presence or absence of light was largely irrelevant to the rats in the common-cause group but highly relevant to rats in the causal-chain group. Our hypothesis explains why.

According to our RR hypothesis, nonhuman animals do not possess the ability to access relations in an omnidirectional fashion (Halford et al., 1998): that is, nonhuman causal representations are purely unidirectional; they can access the effect given the cause but not the cause given the effect[2]. Thus, according to our RR hypothesis—and contrary to the single effect learning model—rats in the common-cause *Observe* group did *not* reason diagnostically when observing tone. Instead, they simply expected food via second-order conditioning. The presence or absence of light was highly relevant to rats in the causal-chain group, on the other hand, because the explicit absence of the causal node (light) in a causal relation (light -> food) inhibits inferring the effect (see again Blaisdell et al., in press).

Our hypothesis leads to a testable prediction. According to our RR hypothesis, increasing the salience of the common cause (e.g., by replacing the light with a blaring tone or footshock) should *increase* the expectation of food in the common-cause group in the *Observe* condition since second-order conditioning will be stronger. In contrast, the single-effect learning model predicts that increasing the salience of the common cause will *decrease* the expectation of food in the common-cause *Observe* condition because rats will be less likely to infer food if the common causal event is harder to overlook or forget about.

Dwyer et al. (forthcoming) have recently replicated Blaisdell et al.'s (2006) experiment and proposed an alternative, purely associative account of the rats' behavior. Dwyer et al. found that lever presses and nose poking were inversely correlated regardless of training condition. Thus, they suggest that nose poking is reduced in the *Intervene* condition relative to the *Observe* common-cause condition through a simple effect of "response competition."

Dwyer et al.'s (forthcoming) results are consistent with one part of our hypothesis: i.e., that interventions interfere with non-causal associations. But Dwyer et al.'s hypothesis does not account for the fact that there was no significant difference in nose poking between the *Intervene* and *Observe* groups in the direct cause condition of Blaisdell et al.'s (2006) and Leising et al.'s (2008) results. Thus response competition alone cannot account for the rats' behavior. Our RR hypothesis can. Our hypothesis

posits that the competition between lever pressing and nose poking varies inversely with the causal power of the relation. When causal power is high, interventions do not interfere and rats draw the appropriate causal conclusions.

Our representational-level explanation of the rats' behavior only relies on computational abilities for which there is solid evidence with rats: i.e., associative integration (e.g., sensory preconditioning and second-order conditioning), encoding and integration of inter-event intervals (i.e., temporal maps), surprise at the omission of anticipated events (see discussion in Blaisdell et al., in press), and functionally compositional first-order relational representations (Penn et al., 2008a). Thus, our representational-level hypothesis is 1) more consistent with what we know about rats' representational capabilities and limitations, and 2) more consistent with the behavioral results reported by Blaisdell et al. (2006), than is a purely associative or a purely rational model.

It is important to emphasize, however, that our representational-level hypothesis does not replace or compete with the single-effect learning model. The single-effect learning model provides a cogent, minimal rational model of the rats' behavior. Our representational-level hypothesis is an explanation for *how* a particular organism approximates the behavior predicted by this minimal rational model given a particular set of representational-level constraints (e.g., no higher-order relational integration). Without a minimal rational model, we would have no idea what function the rats are approximating or why this is rational. Without a plausible representational-level hypothesis, we would have no idea *how* they are approximating that model or why their behavior sometimes deviates from the rational model. In our view, comparative cognitive psychology needs both kinds of models—but it needs to keep them distinct.

4   Do Chimpanzees Have a Theory of Mind (ToM)?

Premack and Woodruff (1978) coined the term, "Theory of Mind" (ToM), to refer to the ability to explain and predict others' behavior by reasoning about the causal role played by mental states such as perceptions, intentions, goals, and beliefs. Needless to say, Premack and Woodruff's (1978) seminal question, "Does the chimpanzee a Theory

of Mind?", has spurred an enormous and contentious literature (for recent reviews see Call, 2007; Call & Tomasello, 2008; Penn & Povinelli, 2007b, in press).

From the very outset, Premack and Woodruff's (1978) question entailed both a *functional-level* explananda—i.e., "Do chimpanzees act *as if* they understand that others have their own perceptions, goals and beliefs?"—and a *representational-level* explananda—i.e., "Do representations of others' unobservable mental states play a causal role in chimpanzee social cognition?" Although Dennett (1978) and many other philosophers quickly pointed out that these two explananda require different kinds of evidence, comparative researchers have largely glossed over the fundamental distinction between these two levels of explanation for the past quarter-century (see also Dennett, 1987; Heyes, 1998; Povinelli, 1999; Povinelli et al., 2003).

At present, the prevailing consensus among comparative researchers is that chimpanzees possess a rudimentary version of a ToM in which functionally individuated representations of others' psychological states play an inferentially coherent and causally efficacious role (Call & Tomasello, 2008; Suddendorf & Whiten, 2003; Tomasello & Call, 2006). Some researchers extend the same claims to monkeys, scrub jays and dolphins (see, for example, Emery & Clayton, in press; Herman, 2006; Santos et al., 2007). Call and Tomasello (2008) sum up the prevailing consensus as follows:

> All of the evidence reviewed here suggests that chimpanzees understand both the goals and intentions of others as well as the perception and knowledge of others. Moreover, they understand how these psychological states work together to produce intentional action; that is, they understand others in terms of a relatively coherent perception–goal psychology in which the other acts in a certain way because she perceives the world in a certain way and has certain goals of how she wants the world to be... In a broad construal of the phrase 'theory of mind', then, the answer to Premack and Woodruff's pregnant question of 30 years ago is a definite yes, chimpanzees do have a theory of mind" (p. 191).

We have long argued that this consensus position is mistaken (e.g., Penn & Povinelli, 2007b; Penn & Povinelli, in press; Povinelli, 1999; Povinelli et al., 2000; Povinelli & Giambrone, 1999; Povinelli & Vonk, 2003, 2004). But so far we have failed to make much of a dent in its popularity. So in the present section, we make our point in a

somewhat novel fashion. Just as we did in the case of causal reasoning above, we argue below that comparative researchers are confusing functional- and representational-level claims and have failed to seek out a minimal rational model of chimpanzees' ToM-like behavior. Then we propose a representational-level explanation for how chimpanzees approximate a ToM without actually having one.

4.1   The ToM-Without-Beliefs Hypothesis

To make our case, we will focus on a single seminal set of experiments conducted by Hare et al. (2000; 2001) which is widely considered to be the "breakthrough" evidence in the comparative ToM debate (Tomasello et al., 2003a).

In Hare et al.'s (2001) protocol, two chimpanzees—one subordinate to the other—were kept in separate chambers on either side of a middle area. Two cloth bags in the middle chamber served as hiding places for small food items. Opaque doors on each side chamber prevented the respective chimpanzees from entering the middle chamber and retrieving the food until the doors were raised. On each trial, the subordinate's door was partially raised while the food was being hidden, allowing the subordinate to peek out and see where the food items were placed and whether or not the dominant was peering out from his own chamber. On each trial, the dominant's door was either partially raised or completely closed while the food items were placed in one of the two containers. Once the food had been placed, the dominant's door was closed and the subordinate was released into the middle chamber before the dominant was released as well.

Hare et al. (2001) reported a number of experimental conditions based on this protocol. In only one of these experiments, however, was the critical metric statistically significant[3]. In the *Uninformed* condition of Experiment 1, the dominant's door was kept closed while the food was hidden and the subordinate could see that the dominant's door was closed; in the control condition, the dominant could see where the reward was hidden and the subordinate could see that the dominant was watching. The subordinate "approached" the hidden food more often in the *Uninformed* condition than in the control condition.

On the basis of this result, Hare et al. (2001) concluded that "chimpanzees know what individual groupmates do and do not know, that is, what individual groupmates have and have not seen in the immediate past" (p.148). Tomasello, Call and Hare (2003a) go on to cite these experiments as "breakthrough" (p.154) evidence that chimpanzees "understand some psychological states in others" (p. 156).

Importantly, these researchers are not claiming that chimpanzees have a full-blown, human-like theory of mind. Tomasello et al. (2003a) admit that "there is no evidence anywhere that chimpanzees understand the beliefs of others (see also Call & Tomasello, 2008). Since the ability to represent and reason about contentful, epistemic representational states was the sine qua non of having a "Theory of Mind" in Premack and Woodruff's (1978) original sense of the term, Call and Tomasello are clearly arguing for a novel and, in our opinion, incoherent construal of what it means to have a ToM. But let's put aside for the moment the question of whether it makes any sense to use the term, "Theory of Mind", to refer to a cognitive system that does not have any representation of mental states qua representational states (but see Penn & Povinelli, 2007b; Penn & Povinelli, in press). For the purposes of the present chapter, we will refer to the original hypothesis formulated by Premack and Woodruff (1978) as the "ToM" hypothesis and the position currently favored by Call, Tomasello and many other comparative researchers as the "ToM-Without-Beliefs" hypothesis.

4.2   Is It Rational to Claim that Chimpanzees Have a ToM Without Beliefs?

On the one hand, Call and Tomasello (2008) acknowledge that chimpanzees do *not* "appreciate that others have mental representations of the world that drive their actions" (p. 191). Yet they nevertheless claim that chimpanzees "understand both the goals and intentions of others as well as the perception and knowledge of others" (p. 191).

What does it mean to say that an animal understands another's "goals" as intentional, psychological states yet does not understand that others have mental representations that drive their actions? As far as we can tell, this incongruous claim only makes sense if it is taken on an *as if* basis: i.e., chimpanzees act *as if* they understand that others have goals but not beliefs.

Of course, nobody disputes that animals of many taxa, not just chimpanzees, act in ways that accord well with the predictions of our commonsense folk psychology or what Dennett (1987) called the "Intentional Stance." With respect to Hare et al.'s (2000; 2001) results, for example, there has never been any dispute about the fact that chimpanzees act *as if* they understand that others can see things (Povinelli, 1999). One might even argue that a folk psychological explanation provides a "rational" (albeit informal and ill-specified) model of the chimpanzees' behavior in the sense that a social animal without any representational or computational constraints would likely increase both its proximate and inclusive fitness if it reasoned in terms of its rivals' internal mental states. But as Dennett (1978) first pointed out thirty years ago, when this whole debate was just getting started, claiming that an animal acts *as if* it understands others' psychological states is not the same thing as claiming that an animal actually acquires, stores and processes functionally individuated representations of others' psychological states and uses these representations in an inferential and causally efficacious fashion.

To date, there is no evidence that representations of others' mental states are performing any actual causal work in chimpanzee social cognition. Povinelli and Vonk (2003) pointed out that Hare et al.'s (2000; 2001) results could be parsimoniously explained by postulating that the subordinate chimps were reasoning solely about the *observable* behavior of their rivals rather than their rivals' *unobservable* psychological states. For example, the behavior of the subordinates might result from a simple strategy glossed by *<Don't go after food if a dominant who is present has oriented towards it>*. The additional claim that the chimpanzees adopted this strategy because they understood that *<The dominant knows where the food is located>* may be intuitively appealing but it is causally superfluous.

Likewise, although there is abundant evidence that apes and monkeys act *as if* they are taking the visual perspective of others into account (e.g., Flombaum & Santos, 2005; Hare et al., 2006), there is no evidence that they are actually representing or reasoning about others' subjective visual experience as distinct from the observable behavioral cues causally related to others' actions in the world (Penn & Povinelli, in press). Nor is there any evidence that nonhuman primates understand that others have a subjective visual experience analogous to their own (Povinelli et al., 2000).

All of the evidence collected to date suggests that chimpanzees only represent others' goals and intentions in terms of *external* states of the environment and *observable* behavioral cues but do not understand that others have *internal* mental representations of goals and *unobservable* intentions which causally guide others' behavior (cf. Tomasello et al., 2005). In other words, all of the existing evidence is consistent with the hypothesis that chimpanzees act *as if* others are goal-directed agents but do not actually understand that others have mental representations that drive their actions (see reviews in Penn & Povinelli, 2007b; Penn & Povinelli, in press).

4.3 The Behavioral Abstraction Hypothesis as a Minimal Rational Model of Social Cognition in Chimpanzees

Based on the lack of any evidence that representations of others' mental states play a causal role in chimpanzee social cognition, Povinelli and Vonk (2003, 2004) postulated the "Behavioral Abstraction" (BA) hypothesis. Just as the single-effect learning model proposed by Waldmann et al. (2008) posits a simpler and more plausible rational model of causal learning than a causal Bayes net alternative, herein we will argue that Povinelli and Vonk's BA hypothesis proposes a simpler and more plausible rational model of nonhuman social cognition than the ToM-Without-Beliefs hypothesis favored by Call and Tomasello (2008).

According to the BA hypothesis, chimpanzees (and humans as well) possess a psychological system, S_b, composed of three components:

1. a database of representations of both specific behaviors and statistical patterns of behaviors abstracted across multiple instances of specific behaviors and specific individuals (these representations may be formed either by direct experience and/or may be epigenetically canalized);

2. a network of statistical relationships that adhere between and among the specific behaviors and invariants in the database;

3. an ability to use these representations and statistical regularities to compute the likelihood of others' specific future actions.

For example, in reasoning about the goal-directed behavior of other animate agents, the BA hypothesis postulates that chimpanzees act *as if* they learn abstract rules about the general behavioral patterns of their conspecifics (e.g., *<others who look hungry and who are oriented towards a piece of food are likely to try and get that food>*) as well as concrete representations about the past and present behavior of particular conspecifics (e.g., *<Abe looks hungry>*, *<A few minutes ago, Abe turned his face and eyes towards that piece of food>*). Chimpanzee then reason about the future behavior of others using this "database" of representations and rules.

The BA hypothesis postulates that only humans have an *additional* psychological system, S_{b+ms}, that uses the representations in S_b to form higher-order representations of unobservable mental states and causal relations involving those mental states (e.g., *<others who look hungry are* feeling *hungry>*, *<others who* feel *hungry are likely to try and get food>*, *<others who orient their head and eyes towards something* see *what they are looking at and* know *it is there>*, *<Abe is* feeling *hungry>*, *<Abe* knows *where the food is located>*). The ToM-Without-Beliefs hypothesis, on the other hand, posits that chimpanzees as well as humans form representations of others' psychological states and that these representations of others' psychological states play a causal role in chimpanzees' social inferences over and above chimpanzees' representations of others' observable behavior (see again Call & Tomasello, 2008). This is the crucial difference between the BA hypothesis and the ToM-Without-Beliefs hypothesis.

Importantly, the BA hypothesis and the ToM-Without-Beliefs hypothesis do *not* differ in terms of the subject's learning mechanisms, powers of reasoning or degree of inferential flexibility (the BA hypothesis is more generous than the RR hypothesis, as we will see below). The BA hypothesis simply postulates that nonhuman animals do not reason about unobservable mental states or causal relations involving those mental states without specifying *why* chimpanzees lack these abilities. Critics who claim that the BA hypothesis limits chimpanzees to "mindless behavioral rules" (Call & Tomasello, 2008) and occurrent stimuli (Tomasello et al., 2003b) are attacking a behaviorist strawman (see again Penn & Povinelli, in press).

Thus, in most social situations, the BA hypothesis and the ToM-Without-Beliefs hypothesis make the same functional predictions. For example, both the BA hypothesis and the ToM-Without-Beliefs hypothesis predict that chimpanzees will avoid competing for food with a dominant rival who was present and oriented when the food was hidden (Hare et al., 2001). Both hypotheses also predict that chimpanzees will differentiate between actors who pretend to be "willing" and those who pretend to be "unwilling" to give them food (Call et al., 2004). The BA hypothesis postulates that chimpanzees reason about these social situations solely on the basis of the observable patterns of behavioral cues—e.g., the pattern of cues indicative of an action being "voluntary" or "involuntary"—without postulating that individuals have an unobservable psychological state causing their behavior. For the BA hypothesis, the distinction between "voluntary" and "involuntary" is akin to the distinction between "in estrus" or "not in estrus": Chimpanzees make this distinction in a flexible and inferentially coherent fashion without thereby positing that females in estrus are "*feeling* fertile." The ToM-Without-Beliefs hypothesis, on the other hand, argues that chimpanzees go beyond the surface behavior to make causal inferences based on the actor's feelings and intentions: i.e., an actor who acts in a "voluntary" fashion has an internal psychological "intention" to perform those actions (Call et al., 2004).

In any situation in which it is not necessary to take an individual's mental state into account in order to predict how that individual will behave—which includes all the experimental protocols cited above—there is no functional difference between these two hypotheses and any empirical debate between the two hypotheses is otiose. Critically, however, the two hypotheses make quite different functional predictions whenever the other individual's representation of the world differs from the focal subject's representation of the world.

As philosophers have long pointed out, a defining characteristic of many mental states—such as beliefs, goals, intentions or perceptions—is that they are *about* something: i.e., they are "Intentional" in a representational sense (Dennett, 1987; Dretske, 1986; Searle, 1983). Furthermore, a defining characteristic of a mental state qua representation is that it can be counterfactual: i.e., it can potentially misrepresent the actual state of affairs. It is obvious that chimpanzees themselves are Intentional subjects.

What is at issue is whether chimpanzees understand that others are Intentional subjects as well. The BA hypothesis predicts that they do not.

The existing evidence seems to strongly support the BA hypothesis. As Call and Tomasello (2008) themselves now admit, "there is currently no experimental evidence that chimpanzees understand false beliefs by, for example, predicting what another will do based on what that other knows (when the subject knows something else to be the case)" (p. 190). Indeed, Call and Tomasello (2008) admit there is not just an absence of evidence; there is evidence of an absence. For example, in the Misinformed condition of Hare et al.'s (2001) experiment, the experimenters initially hid the food while dominant rival was watching and then moved the food to the other hiding location while the dominant's door was down. If the subordinate chimpanzee had been reasoning in terms of its rival's mental states, it should have understood that the rival was misinformed about the location of its goal. In fact, the subordinate chimpanzees did *not* tend to approach the hidden food more frequently than in the control condition—in accordance with the predictions of the BA hypothesis.

Lest there be any worry that our BA hypothesis is unfalsifiable (cf. K. Andrews, 2005; Santos et al., 2007), we have proposed multiple experimental protocols capable of falsifying our hypothesis (see Penn et al., 2008b; Penn & Povinelli, 2007b; Povinelli & Vonk, 2003).

4.4 How To Build a Chimpanzee that Approximates the BA Hypothesis
Given that the BA hypothesis postulates a minimal rational model of chimpanzee social cognition which is more parsimonious, more coherent and more consistent with the empirical evidence than the prevailing mentalistic alternative, let us now turn to the question of *how* a chimpanzee might actually approximate this minimal rational model given what we know about the representational capacities of these animals.

Once again, we employ the representational-level "Relational Reinterpretation" (RR) hypothesis proposed by Penn et al. (2008a): to wit, chimpanzees (like rats and humans) possess the ability to learn and reason about the causal relation between events. But chimpanzees lack the ability to represent higher-order, role-based relations and thus cannot reason about unobservable causal mechanisms or reason by analogy to their own

experience. The question, then, is how can a chimpanzee approximate the rational model set out by the BA hypothesis given its particular representational capabilities and limitations.

We postulate that chimpanzees represent and reason about the *goal-directed* relations that hold between the behavior of animate agents and external states of affairs in the world (see Penn & Povinelli, in press). By *goal-directed* relations we mean the particular causal relationship that animate agents have with objects and states of affairs in the world such that observing an agent's behavioral pattern with respect to an external goal can be used to predict how the agent will act towards that goal in the future. On our RR hypothesis, chimpanzees use representations of concrete goal-directed relations as well as representations of general patterns of goal-directed behavior to predict how others will behave without postulating the existence of *internal* goals (cf. Tomasello et al., 2005). Importantly, when chimpanzees predict how a concrete individual will act, they do not integrate abstract and concrete representations in a structural, role-based or analogical fashion; rather, they match relations on the basis of their perceptual similarity (see Penn et al., 2008a). Thus, chimpanzees reason on the basis of perceptual similarity between a given situation and the situations they have been exposed to in the past; they do not reason in terms of causal mechanisms involving unobservable mental states.

This does not mean that chimpanzees are uninformed statistical learners. As Clark and Thornton (1997) showed, picking out causally relevant *relations* in the world amidst all the salient but spurious correlations presents uninformed statistical learning mechanisms with a computational quagmire. Clark and Thornton (1997) suggest that biological cognizers circumvent the limitations of uninformed statistical learning by employing a range of top-down heuristics, ploys and biases to recognize and reason about the relations that matter. And we agree. Our RR hypothesis postulates that chimpanzees are eminently *relational* reasoners; not just uninformed statistical learners. For example, chimpanzees understand quite a lot about the peculiar causal relation between a competitor's line of sight, the nature of the object being observed, and how the competitor is likely to behave in the near future (Penn & Povinelli, in press; Povinelli et al., 2000). This relational inference does not require mentalistic or analogical reasoning; but it is certainly no mean cognitive feat and far exceeds the capability of a purely

associative learner. Via innate mechanisms as well as ontogenetically canalized learning, chimpanzees (as well, presumably, as many other nonhuman species) come to possess a variety of heuristics, ploys and biases for picking out the causally relevant features of other agents' goal-directed relationships and for reasoning about others' behavior in a relational fashion. In this way, chimpanzees *approximate* having a ToM well enough to fool the average comparative researcher.

5 How to Be Comparatively Rational about Animal Cognition

We have argued that there is a common mistake being committed by comparative psychologists studying ToM and causal cognition in nonhuman animals. In both cases, the explananda driving comparative research have been framed in ways that conflate functional- and representational-level considerations. There is, we believe, a much better (and more rational) way to approach comparative explananda:

First, functional-level claims must be clearly distinguished from representational-level claims. For example, whether a given organism behaves in a way that is consistent with ascriptions of 2nd order intentionality—e.g., "chimpanzees know what others do and do not know"—is *not* the same question as whether a given organism forms 2nd order representations of another organism's mental states (Dennett, 1987). Whether a given organism behaves in a way that approximates a given rational model of causal reasoning is *not* the same question as whether a given organism actually represents and reasons about the entities, variables and relationships posited by that model.

Second, in comparing and evaluating functional-level models, it is necessary to distinguish between normative rational models that are unconstrained by any concerns for computational feasibility or biological plausibility and "minimal" rational models that provide more parsimonious, less computationally demanding and/or more biologically plausible accounts of the function the cognizer is computing (Waldmann et al., 2008). Both kinds of rational models have a role to play in cognitive science. But the minimal rational model provides the stronger bridge towards a representational-level understanding of the computations the animal is actually performing.

Third, once a minimal rational model for a given behavior has been defined, comparative researchers should develop representational-level hypotheses about *how* a particular organism actually approximates these rational models given the representational constraints under which it is operating. A representational-level hypothesis should also explain when and why a given organism deviates from the predictions of a rational model. This does not negate the value of the rational model. Rational models tell us why an organism is behaving the way it is. Representational-level models tell us *how* the organism approximates those rational norms and *why* it sometimes goes wrong.

In sum, we believe that both human and nonhuman animals have evolved to be *approximately* rational. We believe it is high time that comparative psychologists become approximately rational as well.
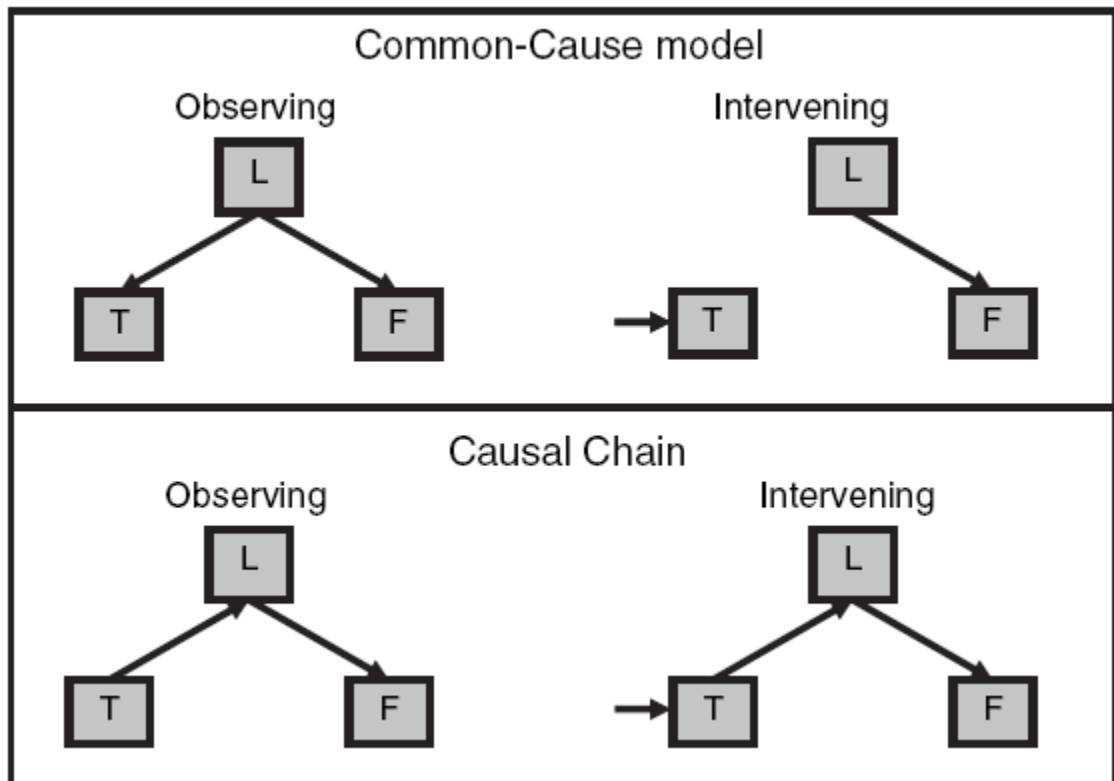

Acknowledgments

Figures

Figure 1

Common-cause and causal-chain models from Blaisdell et al. (2006) where *L* is Light, *T* is Tone and *F* is the delivery of food. The left diagram in each group represents the causal relations in the *Observe* groups. The right diagram in each group represents the causal relations in the *Intervene* groups.

References

Anderson, J. R. (1990) *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Andrews, G., & Halford, G. S. (2002) A cognitive complexity metric applied to cognitive development. *Cognitive Psychology 45*(2), 153-219.

Andrews, K. (2005) Chimpanzee Theory of Mind: Looking in All the Wrong Places? *Mind and Language 20*(5), 521-536.

Arcediano, F., Escobar, M., & Miller, R. R. (2005) Bidirectional associations in humans and rats. *Journal of Experimental Psychology: Animal Behavior Processes 31*, 301-318.

Badre, D. (2008) Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. *Trends Cogn Sci 12*(5), 193-200.

Bermudez, J. L. (2003) *Thinking without words*. New York: Oxford University Press.

Blaisdell, A. P., Leising, K. J., Stahlman, W. D., & Waldmann, M. R. (in press) Rats Distinguish between Absence and Lack of Information. *International Journal of Comparative Psychology*.

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

Buehner, M. J., & Cheng, P. W. (2005) Causal Reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning*. New York: Cambridge University Press.

Bunge, S. A., Wendelken, C., Badre, D., & Wagner, A. D. (2005) Analogical reasoning and prefrontal cortex: evidence for separable retrieval and integration mechanisms. *Cereb Cortex 15*(3), 239-249.

Call, J. (2007) Past and present challenges in theory of mind research in nonhuman primates. *Prog Brain Res 164*, 341-353.

Call, J., Hare, B., Carpenter, M., & Tomasello, M. (2004) 'Unwilling' versus 'unable': chimpanzees' understanding of human intentional action. *Developmental Science 7*(4), 488-498.

Call, J., & Tomasello, M. (2008) Does the chimpanzee have a theory of mind? 30 years later. *Trends Cogn Sci 12*(5), 187-192.

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

Christoff, K., & Keramatian, K. (2007) Abstraction of mental representations: Theoretical considerations and neuroscientific evidence. In S. A. Bunge & J. Wallis (Eds.), *Perspectives on Rule-Guided Behavior*: Oxford University Press.

Clark, A., & Thornton, C. (1997) Trading Spaces: Computation, Representation, and the Limits of Uninformed Learning. *Behavioral and Brain Sciences 20*, 57-90.

Dennett, D. C. (1978) Beliefs about beliefs. *Behavioral and Brain Sciences 4*, 568-570.

Dennett, D. C. (1987) *The intentional stance*. Cambridge, Mass.: MIT Press.

Doumas, L., & Hummel, J. E. (2005) Approaches to Modeling Human Mental Representations: What Works, What Doesn't and Why. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning*. New York: Cambridge University Press.

Dretske, F. I. (1986) Misrepresentation. In R. Bogdan (Ed.), *Belief: Form, Content and Function* (pp. 17-36). New York: Oxford University Press.

Dwyer, D. M., Starns, J., & Honey, R. C. (forthcoming) 'Causal Reasoning' in Rats: A Reappraisal. *Journal of Experimental Psychology: Animal Behavior Processes*.

Emery, N. J., & Clayton, N. S. (in press) How to build a scrub-jay that reads minds. In S. Itakura & K. Fujita (Eds.), *Origins of the social mind: evolutionary and developmental views*. Tokyo: Springer Japan.

Flombaum, J. I., & Santos, L. R. (2005) Rhesus monkeys attribute perceptions to others. *Current Biology 15*(5), 447-452.

Fodor, J. A., & Pylyshyn, Z. W. (1988) Connectionism and Cognitive Architecture. *Cognition 28*, 3-71.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004) A Theory of Causal learning in children: causal maps and bayes nets. *Psychological Review 111*(1), 1-31.

Hadley, R. F. (1997) Cognition, Systematicity and Nomic Necessity. *Mind and Language 12*(2), 137-153.

Halford, G. S., Andrews, G., Dalton, C., Boag, C., & Zielinski, T. (2002) Young Children's Performance on the Balance Scale: The Influence of Relational Complexity. *Journal of Experimental Child Psychology 81*(4), 417-445.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998) Processing capacity defined by relational complexity: implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences 21*(6), 803-831; discussion 831-864.

Hare, B., Call, J., Agnetta, B., & Tomasello, M. (2000) Chimpanzees know what conspecifics do and do not see. *Animal Behaviour 59*(4), 771-785.

Hare, B., Call, J., & Tomasello, M. (2001) Do chimpanzees know what conspecifics know? *Animal Behaviour 61*(1), 771-785.

Hare, B., Call, J., & Tomasello, M. (2006) Chimpanzees deceive a human competitor by hiding. *Cognition 101*(3), 495-514.

Herman, L. M. (2006) Intelligence and rational behavior in the bottlenosed dolphin. In S. Hurley & M. Nudds (Eds.), *Rational Animals?* Oxford: Oxford University Press.

Heyes, C. M. (1998) Theory of mind in nonhuman primates. *Behavioral and Brain Sciences 21*(1), 101-114; discussion 115-148.

Holyoak, K. J., & Hummel, J. E. (2000) The proper treatment of symbols in a connectionist architecture. In E. Dietrich & A. B. Markman (Eds.), *Cognitive Dynamics: Conceptual change in humans and machines* (pp. 229-263). Mahwah, NJ: Erlbaum.

Hummel, J. E., & Holyoak, K. J. (1997) Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review 104*, 427-466.

Hummel, J. E., & Holyoak, K. J. (2005) Relational Reasoning in a neurally plausible cognitive architecture. *Current Directions in Psychological Science 14*(3), 153-157.

Kacelnik, A. (2006) Meanings of Rationality. In S. Hurley & M. Nudds (Eds.), *Rational Animals?* Oxford: Oxford University Press.

Karin-D'Arcy, M. R., & Povinelli, D. J. (2002) Do Chimpanzees know what each other see? a closer look. *International Journal of Comparative Psychology 15*, 21-54.

Kroger, J. K., Holyoak, K. J., & Hummel, J. E. (2004) Varieties of sameness: the impact of relational complexity on perceptual comparisons. *Cognitive Science 28*(3), 335-358.

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

Marr, D. (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W. H. Freeman.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008a) Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences 31*(2), 109-178.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008b) Darwin's triumph: Explaining the uniqueness of the human mind without a deus ex machina. *Behavioral and Brain Sciences 30*(2).

Penn, D. C., & Povinelli, D. J. (2007a) Causal cognition in human and nonhuman animals: A comparative, critical Review. *Annual Review of Psychology 58*, 97-118.

Penn, D. C., & Povinelli, D. J. (2007b) On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. *Philosophical Transactions of the Royal Society B 362*, 731-744.

Penn, D. C., & Povinelli, D. J. (in press) The Comparative Delusion: the 'behavioristic'/'mentalistic' dichotomy in comparative Theory of Mind research. In R. Samuels & S. P. Stich (Eds.), *Oxford Handbook of Philosophy and Cognitive Scence*. Oxford: Oxford University Press.

Povinelli, D. J. (1999) Social understanding in chimpanzees: New evidence from a longitudinal approach. In P. Zelazo, J. Astington & D. Olson (Eds.), *Developing theories of intention: Social understanding and self-control* (pp. 195-225). Hillsdale, NJ: Erlbaum.

Povinelli, D. J. (2000) *Folk Physics for Apes: the chimpanzee's theory of how the world works*. Oxford: Oxford University Press.

Povinelli, D. J., Bering, J., & Giambrone, S. (2003) Chimpanzee 'pointing': Another error in the argument by analogy? In S. Kita (Ed.), *Pointing: Where language, culture and cognition meet*. Hillsdale, NJ: Erlbaum.

Povinelli, D. J., Bering, J. M., & Giambrone, S. (2000) Toward a Science of Other Minds: Escaping the Argument by Analogy. *Cognitive Science 24*(3), 509-541.

Povinelli, D. J., & Giambrone, S. (1999) Inferring other minds: Flaws in the argument by analogy. *Philosophical Topics 27*, 167-201.

Povinelli, D. J., & Vonk, J. (2003) Chimpanzee minds: suspiciously human? *Trends in Cognitive Sciences 7*(4), 157-160.

Povinelli, D. J., & Vonk, J. (2004) We Don't Need a Microscope to Explore the Chimpanzee's Mind. *Mind and Language 19*(1), 1-28.

Premack, D., & Woodruff, G. (1978) Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences 4*, 515-526.

Santos, L. R., Flombaum, J. I., & Phillips, W. (2007) The Evolution of Human mindreading: how non-human primates can inform social cognitive neuroscience. In S. Platek (Ed.), *Evolutionary Cognitive Neuroscience* (pp. 433-456). Cambridge: MIT Press.

Searle, J. R. (1983) *Intentionality, an essay in the philosophy of mind*. Cambridge [Cambridgeshire] ; New York: Cambridge University Press.

Shanks, D. R. (2005) Connectionist models of basic human learning processes. In G. Houghton (Ed.), *Connectionist models in cognitive psychology* (pp. 45-82). Hove, East Sussex: Psychology Press.

Shannon, C. E., & Weaver, W. (1949) *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Shettleworth, S. J. (1998) *Cognition, Evolution and Behavior*. New York: Oxford University Press.

Smolensky, P. (1991) Connectionism, constituency, and the language of thought. In B. Loewer & G. Rey (Eds.), *Meaning in Mind: Fodor and his Critics* (pp. 201-227). Oxford: Basil Blackwell.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003) Inferring causal networks from observations and interventions. *Cognitive Science 27*(3), 453-489.

Suddendorf, T., & Whiten, A. (2003) Reinterpreting the Mentality of Apes. In J. Fitness & K. Sterelny (Eds.), *From Mating to Mentality: Evaluating Evolutionary Psychology* (pp. 173-196): Psychology Press.

Tomasello, M., & Call, J. (2006) Do chimpanzees know what others see-- or only what they are looking at? In S. Hurley & M. Nudds (Eds.), *Rational Animals?* Oxford: Oxford University Press.

Tomasello, M., Call, J., & Hare, B. (2003a) Chimpanzees understand psychological states - the question is which ones and to what extent. *Trends in Cognitive Sciences 7*(4), 153-156.

Tomasello, M., Call, J., & Hare, B. (2003b) Chimpanzees versus humans: it's not that simple. *Trends in Cognitive Sciences 7*(6), 239-240.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005) Understanding and sharing intentions: the origins of cultural cognition. *Behavioral and Brain Sciences 28*, 675-691.

van Gelder, T. (1998) The Dynamical Hypothesis in Cognitive Science. *Behavioral and Brain Sciences 21*(5), 615-628.

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

Yin, H., Barnet, R. C., & Miller, R. R. (1994) Second-order conditioning and Pavlovian conditioned inhibition: operational similarities and differences. *J Exp Psychol Anim Behav Process 20*(4), 419-428.

Notes

[1] Numerous researchers and philosophers have raised concerns about the nature of "causal efficacy" in a nonclassical computational system (e.g., Smolensky, 1991). This is not the forum to tackle this thorny issue directly (but see Hadley, 1997; Hummel & Holyoak, 1997). In our sense, a representational-level explanation simply is any explanation that claims to specify how information is represented by the causally efficacious constituents of a cognitive process. If there are no such constituents, as some believe (van Gelder, 1998), then there is no representational-level explanation possible.

[2] It appears that rats may form bidirectional associations under certain training conditions (see Arcediano et al., 2005). However, our hypothesis is that rats' ability to retrieve a representation of an antecedent event works through purely associative mechanisms and does not take into consideration the causal power of the relation. Omnidirectional access is necessary for true diagnostic reasoning since the causal power of the relation is integral to the likelihood that the candidate cause is the true cause of an observed effect.

[3] Hare et al. (2000; 2001) used two metrics—'retrieve and 'approach'—to measure the animals' performance on these tests. The first recorded the percentage of food items actually retained by the subordinate. The second recorded the percentage of trials on which the subordinate left its own chamber and crossed into the middle chamber prior to the dominant being released. As Karin-D'Arcy and Povinelli (2002) note, given the fact that the dominant chimp often did not know where the food was located and given the fact that the subordinate was given a sizeable headstart, it is hardly meaningful that the subordinate retrieved more food.