Computational Comparison of Children and Apes on a Non-Verbal False Belief Task

Margeaux F. Ciraolo (ciraolo@hawaii.edu)[1]
Samantha M. O'Hanlon (sohanlon@hawaii.edu)[1]
Leonidas A. A. Doumas (alex.doumas@ed.ac.uk)[1,2]
[1] Department of Psychology, University of Hawai'i at Mānoa, Honolulu, HI, USA
[2] School of Philosophy, Psychology and Language Sciences; University of Edinburgh; Edinburgh, UK

Abstract

The key difference between the cognitive abilities of humans and other animals may be the ability to reason relationally; models of relational reasoning are one way to demonstrate this proposed difference. The present work uses the DORA model to simulate a task designed to assess the theory of mind capabilities of 4- and 5-year-old children and apes. In the original experiment, the apes and children successfully completed a number of control tasks in which they used cues from an experimenter to reason about the hiding location of a reward. However, only the children succeeded on the critical manipulation in which it was necessary to infer what the experimenter knew. The simulations presented herein demonstrate that the apes' performances across all tasks can be accounted for by simple rule use. Conversely, the 4-year-olds succeeded via relational inference and learning; 5-year-olds alone had the requisite relational structures predicated beforehand.

Keywords: relational reasoning; cognitive development; computational models; comparative cognition.

Introduction

There is a long tradition in behavioral research of comparing the cognitive achievements of humans and non-human animals in order to assess the similarities and differences in their thinking and reasoning. Interestingly, many of the cognitive capabilities once thought to be the most "human" (e.g., transitive inference, language, relational thinking, hierarchical reasoning, mental state attribution) are claimed to have been observed in various animal species (see e.g., Bergman, Beehner, Cheney, & Seyfarth, 2003; Cook & Wasserman, 2007; Dally, Emery, & Clayton, 2006; Gentner, Fenn, & Margoliash, 2006; Lazareva et al., 2004). However, despite the striking similarities in the abilities of human and non-human animals, human cognition remains singular in the animal kingdom. Specifically, human cognition appears to possess a flexibility not observed in other animals (see below). The differences between humans' and nonhuman animals' cognition raise two very important questions. First, what are the cognitive processes that allow for the behavioral flexibility observed in human reasoning but not in that of other species? Second, to what extent are the mechanisms underlying these processes shared across species?

In an attempt to address the first question, Penn, Holyoak, and Povinelli (2008) suggest that the cognitive process that lies at the heart of the observed differences between human and nonhuman animals is relational reasoning. Relational reasoning, in short, is reasoning about some object based on the role that it plays rather than its physical features alone. Penn et al. argue that this ability underlies the flexibility and structural sensitivity required for many uniquely human capabilities (e.g., language production, art, science, and mathematics; see also, Medin, Goldstone, & Gentner, 1993). In order to reason relationally, a system must be able to represent relations as explicit entities that can be dynamically bound to arguments (i.e., they must be *predicated*; Doumas & Hummel, 2005). Penn et al. (2008) argue that it is this precise capacity that differentiates human and non-human cognition, and they make a strong argument that, thus far, there is insufficient evidence to conclude that any non-human species possess this ability.

The current research attempts to address the second question within the context of a non-verbal false belief task (i.e., that of Call & Tomasello, 1999). A series of computer simulations utilizing the DORA model of human relational learning (Doumas, Hummel, & Sandhofer, 2008) was conducted in order to demonstrate that the capacity for relational reasoning can explain the differences in task performance between apes and human children, as well as the developmental trends observed in 4- and 5-year-old children.

Methods

The following sections describe the non-verbal false belief tasks that were given to children and apes, a brief description of the LISA/DORA models of relational reasoning (for a more thorough explanation, see Doumas et al., 2008), and how the DORA model was used to simulate the behavioral data collected by Call and Tomasello (1999).

Task Description

Theory of mind tasks can be understood as relational in nature because they require a subject to reason about the mental contents of another, which involves using a higher-order relational structure to cast a belief state on some proposition (Penn & Povinelli, 2007). [1] Theory of mind is a hotly debated topic within the comparative literature (see

[1] For example, *knows*(communicator, *contains*(box, reward)). Words in italics represent predicated relational concepts (i.e., abstracted relations that are independent of the objects to which they are bound). The objects within the parentheses denote the actors fulfilling these roles. In this example: box (actor; the object doing the containing) contains (predicated relational concept) reward (patient; the object being contained). *Knows* is another predicated relational concept that is taking the *contains*(box, reward) predicate as the patient (the thing that is known about), thereby forming a higher-order relational structure.

Penn & Povenilli, 2007, for a discussion). Call and Tomasello (1999) demonstrated that 4- and 5-year-old children are capable of reasoning about the false beliefs of an observer (a type of theory of mind task) and attempted to test whether chimpanzees (*Pan troglodytes*) and orangutans (*Pongo pygmaeus*) also possess this ability.

The original experiment was separated into four control tasks and one false belief task. In all five tasks, the *hider* would put a reward in one of two boxes, out of view of the participant and in full view of the *communicator*. The communicator would then place a *marker* on the box that she saw the reward go into and the participant had to choose the box that contained the reward. [2] Each control task evaluated participants' abilities to perform an individual component of the false belief task. Task 1 addressed the question of whether the participant was able to choose the marked box in order to obtain a reward. In Task 2, the communicator marked the location of the reward, then the hider moved the reward from one box to the other in full view of the participant; this task assessed the participant's ability to track the movement of the reward when it was visibly displaced. In Task 3, the communicator marked the location of the reward, and then the hider moved the boxes (one of which contained the reward) rather than the reward itself. In Task 4, the hider moved the reward from one box to the other in full view of the participant (as in Task 2) but out of the sight of the communicator before the communicator marked a box; to choose the box containing the reward, the participant had to ignore the communicator's mark when the communicator was known to be wrong (i.e., the participant had seen the reward being placed in the other box).

In the fifth and final manipulation of the task (i.e., the false belief task), understanding what the communicator did and did not know became essential for selecting the box containing the reward. Specifically, the hider switched the locations of the boxes in full view of the participant (as in Task 3) but out of the sight of the communicator before the communicator marked a box (as in Task 4). Of critical importance, when the communicator marked the box in which she had seen the reward hidden, it was the wrong one because she did not know that the boxes had been moved. Thus, Task 5 addressed whether the participant understood that the communicator had been fooled. Tasks 1 - 4 do not require the participant to reason about the mental contents of another while Task 5 requires the participant to recognize that the communicator holds a false belief. All of the apes performed below chance on these false belief trials (Task 5), despite the fact that they performed well above chance on the component tasks (Tasks 1 - 4). Thus, although the apes demonstrated that they were capable of choosing the unmarked box when they themselves had seen the reward moved before the communicator marked the box she believed the reward to be in (Task 4), they failed to choose the unmarked box when they had seen the boxes moved before the communicator marked one (Task 5).

[2] The marker was removed after a few seconds when the children were performing the task. In contrast, the marker remained on the box for the apes because performing the task with the marker removed proved too difficult for all but one of the apes, possibly due to working memory constraints (Read, 2008). Otherwise, the tasks were the same for the children and the apes.

Model Description

Although there are many models of relational reasoning (e.g., Falkenhainer, Forbus, & Gentner, 1989; Holyoak & Thagard, 1989), of particular interest to the proposed study are LISA (*Learning and Inference with Schemas and Analogies*), developed by Hummel and Holyoak (1997, 2003) and DORA (*Discovery of Relations by Analogy*), developed by Doumas et al. (2008). Collectively, LISA and DORA account for over 90 phenomena from the human cognitive development literature (e.g., Doumas et al., 2008; Hummel & Holyoak 1997, 2003). DORA was developed from LISA in response to the criticism that the LISA model was not able to account for where the structured representations it uses might originate (Munakata & O'Reilly, 2003; O'Reilly & Busby, 2002; O'Reilly, Bubsy, & Soto, 2003). DORA solves this particular problem by offering a neurally plausible instantiation of how structured representations can be learned from unstructured examples observed in the environment and thus provides an account of how children (and adults) acquire the representations that allow them to reason relationally.

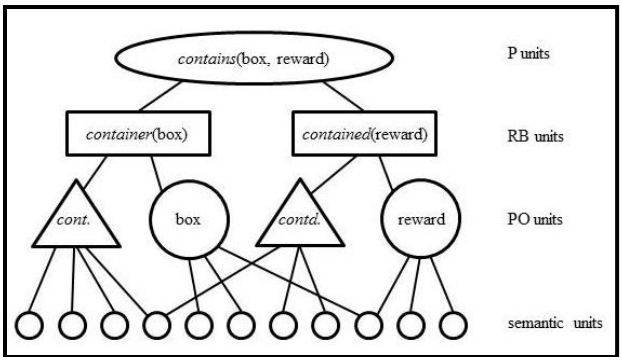

Figure 1. The proposition "box contains reward" is represented at the various levels of localist units. These representations are distributed in the sense that units in the layer above conjunctively code for units in the lower layers.

Nature of Representations Representations within LISA/DORA exist as a hierarchy of distributed and localist units in a layered connectionist architecture (see Figure 1). On the bottom layer of the representational structure are semantic units coding for the features of objects and roles (or predicates) in a distributed fashion. In LISA/DORA, semantic units are shared between predicates and objects for two important reasons. First, it is important for the meaning of some property of an object and the explicit predicate of that property to mean the same thing (Doumas et al., 2008). That is, without a shared pool of semantic units, 'blue' as a feature of the ocean would be unlike 'blue' as a predicate,

which can be cast upon any object (Doumas et al., 2008). Localist predicate-object (or PO) units in the layer above the semantic units act as tokens for individual predicates and objects. Above the PO units, localist role-binding (RB) units link predicate and object units into role-filler pairs. Proposition (P) units in the top layer link sets of RB units together to form whole propositions.

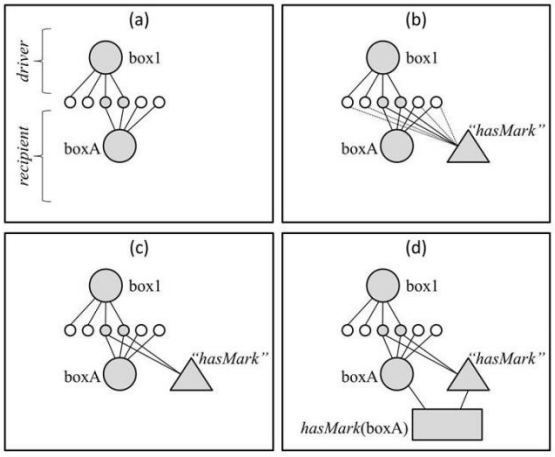

Figure 2. Relational learning in DORA; a new PO unit is recruited that codes for the featural overlap between two objects. This new PO unit becomes an explicit representation of these features.

Flow of Control In LISA/DORA, propositions are divided into two mutually exclusive sets as they enter working memory. The first set, the *driver* (analogous to what the model is "attending to" at any given time; see Figure 2a) controls the sequence of firing events. The patterns of activation imposed on the semantic units by tokens in the driver allow LISA/DORA to retrieve propositions from long term memory into the second set, the *recipient* (analogous to active memory; Cowan, 2001; see Figure 2a), thereby making them available for mapping to propositions in the driver. Activation in the model then flows from the driver into the shared pool of semantic units, which, in turn, causes token units in the recipient to become active (Doumas et al., 2008).

When a proposition becomes active in the driver, role-filler bindings must be represented dynamically on the units that maintain role-filler independence (see, e.g., Hummel & Holyoak, 1997). Unlike LISA (in which binding information is carried via *synchrony* of firing), DORA carries binding information via systematic *asynchrony* of firing. Specifically, roles and the arguments to which they are bound fire in direct sequence as asynchronous couplets. The result is a pattern in which bound role-filler pairs fire in direct sequence and out of synchrony with other bound role-filler pairs. Carrying binding information by *when* units fire allows identity information to be carried independently by *which* units fire. Thus, DORA solves the dynamic binding problem while processing structured symbolic representations in a fundamentally connectionist architecture. The result is a model with representations that include the strengths of both symbolic systems (i.e., structure sensitivity) and connectionist systems (i.e., distributed representations), while suffering the limitations of neither (see, e.g., Doumas & Hummel, 2005, 2010; Hummel & Holyoak, 1997, 2003).

Mapping and Relational Learning Generally speaking, mapping (i.e., the process of comparison) creates opportunities for DORA to predicate new properties. Mappings between units in the driver and the recipient indicate that these units have some properties in common. DORA's mapping algorithm is the same as LISA's; when units are active in both the driver and recipient simultaneously, the model attempts to map them by learning connections between them. Therefore, as units in the driver become active (i.e., as DORA 'thinks' about them), they will activate structurally and semantically similar units in the recipient through any shared semantic feature units. As a consequence, DORA maps structurally and semantically similar propositions across the driver and the recipient.

In DORA, learning is a function of the ability to compare (see Doumas et al., 2008 for details). DORA begins with simple feature vector representations of objects (i.e., a PO unit connected to semantic units). As DORA goes through the process of comparing two objects, and they become co-active, the corresponding features of those objects also fire in unison. Any semantic units that the two objects have in common become highlighted by virtue of receiving twice the activation that unshared units receive (see Figure 2a). DORA then recruits a new PO unit and learns connections between said unit and the active semantic units in proportion to their activation via a Hebbian learning rule (i.e., stronger connections to more active units; see Figure 2b). This process generates an explicit representation of *all* of the properties shared between the two objects, including those which may be irrelevant. For example, if a red apple is compared to a red fire engine, the explicit representation of 'red' learning by the model will also carry with it any other features shared by the compared objects (e.g., both objects also might also contain the feature 'shiny'). Consequently, additional examples of 'red' are needed in order to rid the representation of extraneous features. As DORA compares multiple instances of 'red' objects extraneous features wash out, leaving only the essential features of the concept (see Figure 2c). Doumas et al., (2008) demonstrated how this process, applied over a range of examples, allows DORA to learn explicit structured representations of object properties and relational roles that can be linked together to create complex relational structures (see Figure 2d).

Simulations

The goal of the simulations presented herein is to explain the behavioral data of Call and Tomasello (1999) by demonstrating that the types of relational structures that are available to the reasoner (in this case, the ape or child) influences his/her performance on theory of mind tasks. We accomplish this by manipulating the types of knowledge

structures that are available to the model and comparing its performance to the behavior observed in the original experiment. More specifically, simple knowledge structures may be able to account for nonhuman primate data while more complex (i.e., abstract/relational) knowledge structures may be required to account for the performance of 4- and 5-year old children. That is not to say that humans do not use simpler knowledge structures, but that false belief tasks require these complex structures. The purpose of the present simulation is not to demonstrate how relational representations are acquired; therefore, these structures are built-in from the onset of the simulation. However, importantly, all of the structures we use in the simulation can be easily learned via DORA's predicate learning routines. For a detailed account of how these structures are learned see Doumas et al. (2008).

Simulation 1: Apes To simulate the apes' performance on the non-verbal false belief task, we assumed that, instead of reasoning based upon the actions of an observer, the apes in the study were using selection criteria based on combinations of visual features present in the experimental context. Specifically, if the reward is seen in a particular location, choose that location; otherwise, choose the box with the marker on it. Each task was coded with box1 and box2 objects (represented by PO units); the features of these objects included generic features of boxes, whether they had a mark, and whether the reward was seen being put into them. Two objects representing selection criteria were created and then placed in the driver while the representations of the boxes were loaded into the recipient. The model was then allowed to generate mappings between the selection criteria in the driver and the boxes in the recipient. Whichever selection criterion DORA mapped to in the driver was taken as DORA's "choice" of the box to investigate for the reward. Note that mapping in this manner utilizes only the featural (as opposed to relational) aspects of the task and can therefore be entirely accounted for by associative learning mechanisms. There were four possible mappings DORA could make. A success was counted when the proper selection criterion was placed in correspondence with the proper object; the other three mappings were considered misses. A total of ten trials per manipulation were simulated.

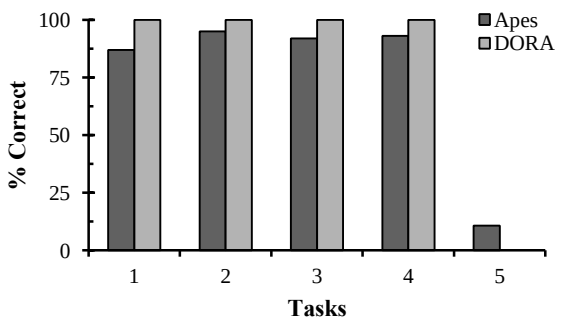

Figure 3. The comparison of the apes' performance to that of the model used in Simulation 1.

The results of Simulation 1 are depicted in Figure 3. The model was 100% accurate on all of the control task simulations (Tasks 1 - 4). In comparison, the apes' scored 87% correct on the task in which they had to choose the marked box (Task 1), 95% correct on the task in which they witnessed the hider move the reward after the communicator had marked the boxes (Task 2), 92% correct on the task in which they witnessed the hider switch the locations of the boxes before the communicator marked one (Task 3), and 93% correct on the task in which they had seen the location of the reward and had to ignore the communicator's (incorrect) mark (Task 4). [3] Similarly, the model performed at 0% on the false belief manipulation (Task 5), while the apes performed at 10.7%. Thus, DORA's performance closely resembled the behavioral data collected from all five of Call and Tomasello's (1999) tasks. For these simulations, we assumed that the model had perfect attention and focused only on the boxes and the task criteria. The small differences that were observed between the model's behavior and the behavioral data could be fit by adding noise into the simulations.

Simulation 2: 4-Year-Old Children Both the children and the apes had little difficulty selecting the location of the reward if they had seen where it had gone, and neither group had difficulty using the marker as a cue for the location of the reward. Simple rule use can easily account for this pattern of behavior, so we assumed that the 4-year-old children were initially reasoning about the task in much the same way as the apes did. However, unlike the apes, children would likely be building and refining more complex representational structures for the task across manipulations and trials.

Therefore, Simulation 2 focused specifically on how the performance of the 4-year-old children changed during the course of the false belief task. This simulation was conducted by first placing the proposition *hasMark*(box) + *select*(box) in the driver and *hasMark*(box1) and *noMark*(box2) in the recipient.[4] Each trial consisted of first allowing DORA to map the representations in the driver to the *hasMark*(box1) and *noMark*(box2) in the recipient, and then use relational inference to select a box. If the model inferred the representation *select*(box2), it was recorded as a hit and all other inferences were counted as misses.[5] Fourteen blocks of four trials each were run for a total of 56 trials. The 4-year-old children and the model both performed below chance on the false belief task and

[3] Due to the unavailability of the experimental data, percentages from tasks 2, 3, and 4 have been estimated from the figures and t-statistics reported by Call and Tomasello (1999).

[4] Here, "+" denotes binding simple representational structures into propositions (i.e., multi-place predicates). See Doumas et al. (2008) for further discussion of how propositions are encoded and used for reasoning in DORA.

[5] The box1 and box2 objects differed only in regard to whether they carried semantic units for being marked or having the reward. Therefore, counter-balancing which box had the 'reward' semantic unit was unnecessary.

exhibited gradual improvement across trials. The 4-year-olds' average performance across all four trials was about 38% correct and the model's was about 41% correct.

According to the changes in DORA's representations across trials, this increase in accuracy is due to the fact that the mark always predicted the presence of the reward; thus DORA's representation of the task initially had the features 'mark' and 'reward' bound to the same PO unit. In regard to the children, this translates to their representations of these features being conflated. Therefore, 4-year-old children would have had to update their conflated representations and generalize a new rule before they could succeed on the false belief task. Likewise, DORA's representation was refined over successive trials (see the above section titled *Relational Learning* for a description of how DORA accounts for this error and its resolution).

Simulation 3: 5-Year-Old Children Seventy-nine percent of the 5-year-old children in Call and Tomasello's (1999) study were successful on their first attempt at the false belief manipulation; we therefore focused on their performance in the first trial and assumed that, by this age, children have built a representation of the rules "If *knows*(x), then *accurate*(x)," and "If *notKnows*(x), then *inaccurate*(x)." On all trials in which the *notKnows*(x) + *inaccurate*(x) proposition was placed into the driver and *notKnows*(communicator) was placed into the recipient, the model was able to infer that, because the communicator did not know the location of the reward, the communicator's mark was inaccurate and, therefore, selecting the marked box was not the correct choice. It is worth noting that the model had perfect attention and task execution, whereas some portion of the children's errors may be attributable to loss of attention and lack of inhibitory control. Together, attention and inhibition are likely explanations for the discrepancy between the model's perfect performance and children's slightly less than perfect performance. Our goal in these simulations was not to adjust parameters unnecessarily (e.g., by adding noise) to more closely fit the data. Instead we were concerned with qualitative fits and making the fewest additional assumptions possible. As such, we did not include properties like reduced attention or noise, but rather sought to simulate general trends using slight variations in the types of knowledge representations available to the model. Specifically, we were able to simulate the behavior of the apes using only holistic feature properties, whereas structured representations were required to simulate the behavior of the children. While these structures can be learned from holistic feature vectors (see Doumas et al., 2008), Penn et al., (2008) argue that it is precisely the capacity to learn and manipulate these structures that differentiates human and non-human cognition.

Conclusion

Call and Tomasello (1999) concluded that the apes were not capable of utilizing the mental contents of the observer to reason successfully on the false belief task. However, the authors provide an alternative explanation, speculating that the task may have been too difficult for the apes, as success would have involved coordinating many different small pieces of evidence. This interpretation does not preclude apes from possessing theory of mind per se; however, there has yet to be a definitive demonstration of theory of mind capabilities in apes (see Penn & Povenelli, 2007 for further discussion).

In support of Call and Tomasello's conclusion, their results were simulated without any information from the observer or the hider. Therefore, it is unlikely that the observer's behavior had any impact on the apes' reasoning. DORA's ability to account for these behavioral data without the structured representations typically thought of as being necessary for relational reasoning suggests that apes succeeded on the control tasks by using simple associative learning alone; namely, retrieving memories of receiving rewards and the associated perceptual features of the task configuration, then mapping those features onto the test configurations. This claim is further substantiated by the apes' failure on the false belief task, in which using relational inference was necessary for reasoning about the mental contents of another.

The results from the simulations of 4- and 5-year-olds provide evidence that children are using relational knowledge (in addition to associative learning mechanisms) to reason about the task. The difference between the performances of 4- and 5-year-olds seems to be whether they possess the particular relational representation required to reason about false beliefs (i.e., *notKnows*(x)). Although these simulations do not provide conclusive evidence that apes lack theory of mind capabilities, they support the notion of relational reasoning being critical to both the observed differences in apes' and humans' performance on false belief tasks and in human and nonhuman animal cognition in general.

References

Bergman, T. J., Beehner, J. C., Cheney, D. L., & Seyfarth, R. M. (2003). Hierarchical classification by rank and kinship in baboons. *Science*, *302*, 1234-1236.

Call, J., & Tomasello, M. (1999). A nonverbal false belief task: The performance of children and great apes. *Child Development*, *70*, 381-395.

Cook, R. G., & Wasserman, E. A. (2007). Learning and transfer of relational matching-to-sample by pigeons. *Psychonomic Bulletin and Review*, *14*, 1107-1114.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87-114.

Dally, J. M., Emery, N. J. & Clayton, N. S. (2006). Food-caching western scrub-jays keep track of who was watching when. *Science*, *312*, 1662-1665.

Doumas, L. A. A., & Hummel, J. E., (2005). Approaches to modeling human mental representations: What works, what doesn't and why. In K. J. Holyoak and R. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning.* Cambridge: Cambridge University Press.

Doumas, L. A. A., & Hummel, J. E., (2010). A computational account of the development of the generalization of shape information. *Cognitive Science*, *34*, 698-712.

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*, 1-43.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure mapping engine: Algorithm and examples. *Artificial Intelligence*, *41*, 1-63.

Gentner, T. Q., Fenn, K.M., Margoliash, D., & Nusbaum, H.C. (2006). Recursive syntactic pattern learning in songbirds. *Science*, *440*, 1204-1207.

Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science*, *13*, 295-355.

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. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*, 220-264.

Lazareva, O. F., Smirnova, A. A., Bagozkaja, M. S., Zorina, Z. A., Rayevsky, V. V., & Wasserman, E. A. (2004). Transitive responding in hooded crows requires linearly ordered stimuli. *Journal of the Experimental Analysis of Behavior*. *82*, 1-19.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*, 254-278.

Munakata, Y., & O'Reilly, R. C. (2003). Developmental and computational neuroscience approaches to cognition: The case of generalization. *Cognitive Studies*, *10*, 76-92.

O'Reilly, R.C., & Bubsy, R. S. (2002) Generalizable relational binding from coarse-coded distributed representations. In T.G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems (NIPS) (Vol. 14)*. Cambridge, MA: MIT Press.

O'Reilly, R. C., Bubsy, R. S., & Soto, R. (2003). Three forms of binding and their neural substrates: Alternative to temporal synchrony. In A. Cleeremans (Ed.), *The unity of consciousness: Binding, integration, and dissociation.* Oxford: Oxford University Press.

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

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

Read, D. W. (2008). Working memory: A cognitive limit to non-human primate recursive thinking prior to hominid evolution. *Evolutionary Psychology*, *6*, 676-714.

Sandhofer, C. M., & Doumas, L. A. A. (2008). Order of presentation effects in learning color categories. *Journal of Cognition and Development*, *9*, 194-221.