UNIVERSITY OF CALIFORNIA
SANTA CRUZ

(DIS)APPEARING MINDS: METHODOLOGICAL ASSUMPTIONS AND EPISTEMOLOGICAL BIASES IN ANIMAL BEHAVIOR AND COGNITION RESEARCH

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

In

PHILOSOPHY

by

Alexis D. Mourenza

December 2016

The Dissertation of Alexis D. Mourenza is approved:

_____
Professor Daniel Guevara, Chair

_____
Professor Nico Orlandi

_____
Professor Paul Roth

_____
Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

Abstract

Alexis D. Mourenza

"(Dis)Appearing Minds: Methodological Assumptions and Epistemological Biases in Animal Behavior and Cognition Research"

A debate in the nonhuman animal cognition literature exists between those providing 'associative' accounts and those providing more 'mentalistic' explanations of the cognitive mechanisms underlying nonhuman animal behavior. The former maintain that all nonhuman animal behavior can be explained as the result of conditioned associations or innate reflexes and the latter propose that some nonhuman animals make use of higher levels of mental representation. One area of inquiry that has received significant attention asks whether humans are unique in their ability to form representations of another's mental states, that is, whether all other animals rely on strictly associative mechanisms that enable a kind of "behavior-reading" rather than "mind-reading." I will propose that the naturalistic tasks put to the subjects allows the formulation of seemingly contradictory conclusions to be drawn from the experimental results. By examining only species-typical traits in a naturalistic setting, researchers are confronted with the possibility that the observed behaviors result from trained associations acquired in the animal subject's individual learning history or stimulus-bound relations acquired in the evolutionary history of the species, rather than resulting from the subject's ability to consciously reason about it's environment (including the other individuals that are part of it).

Richard Samuels and Stephen Stich have proposed that reasoning in humans is underwritten by two distinct neuronal systems, one that is uniquely human and is constituted by a number of higher-order, domain-specific mental modules and another that is shared with other animals and is constituted by a set of cognitively simpler, domain-general problem-solving skills, like association. Ronald Schusterman and colleagues' demonstration of the eventual acquisition of an equivalence concept by their sea lion subject, Rio, in a nonnaturalistic experimental paradigm will be provided as a case-study for the attribution of a dual-processing cognitive system to a nonhuman animal.

Association is purported to be able to account for learning, the structure of mental states, and the way certain thoughts relate to other thoughts. But Eric Mandelbaum has shown that accepting an associative account of one of these mental processes does not entail that we must also accept associative accounts of the others. The primate theory of mind literature will be reexamined in light of a more thorough understanding of the varieties of associationist theses and the conflation of them. I will propose that the burden of proof should be lifted from those who seek to explain numerous diverse complex behaviors in a nonhuman animal species by reference to higher-order cognitive mechanisms when doing so accounts for the larger body of data, rather than advising them to explain a single experimental result in isolation by postulating a complex web of conditioned associations for which no evidence is available to validate the assumption that such associational learning has taken place. Further, ecologically-invalid, nonnaturalistic experimental paradigms eliminate such

assumptions from consideration and thereby provide the strongest evidence of higher-order cognitive mechanisms operating in a nonhuman animal subject.

Chapter 1: Animal Minds

I. Animal minds debate

The 'animal mind,' or 'minded animal' as an object of philosophical inquiry and reflection has a long history. Two distinct approaches are present throughout the history of the debate over whether any nonhuman animal makes use of mental representations when interacting with their environment, or whether their actions are strictly bound by physical stimuli. These approaches, in the case of other animals, are the same as the response to the general 'problem of other minds' as it arises in attributing mental states to other humans: we can argue by analogy or by anecdote. The problem of other minds, whether human or nonhuman, is a problem of access; we cannot directly observe another's mental contents, so all questions asked must be indirectly posed through a (epistemologically constrained) methodological framework and experimental apparatus.

In this chapter I will begin by describing the varieties of evidence available for those wishing to attribute mental states to nonhuman animals and proceed to examine the epistemic virtue of the principle of parsimony as it relates to this area of inquiry. Special attention will be paid to the primary methodological principle of animal behavior and cognition research, i.e. Morgan's Canon, and how it relates to the general principle of parsimony, i.e. Ockham's Razor. I will then provide a punctuated historical overview of the positions that have been defended in the scientific and experimental literature in regards to the question of whether any nonhuman animals possess mental states like those that guide human behavior. I will

close with an outline of the trajectory of the larger project contained in this dissertation.

II. Indirect evidence: analogy and anecdote

An anecdote is a short narrative of an interesting event, often presented with the aim of demonstrating some point to the listener. Anecdotal evidence depends on a form of *abductive reasoning*, i.e. inferring from an observation to a hypothesis that provides the simplest and most likely explanation of that observation. Abductive reasoning is often defined as an 'inference to the best explanation.' An analogical argument, alternatively, is a form of *inductive reasoning* by which an analogy between a model system and a target system are made by means of the presumed similarity between the two. For example, if I am thirsty I may go to the kitchen and pour myself a glass of water. If I see my friend perform such a behavior, I can presume that she was also feeling thirsty. But an analogical argument for the similarity of two systems is only as strong as the reasoning on which the initial similarity was determined, because the claim for further similarity is based on the presumed similarity (i.e. that we are both biological systems that respond to thirst by seeking fluids). Like an anecdotal argument, analogical reasoning is not a deductively valid form of inference, i.e. the truth of the conclusion drawn is not guaranteed by the truth of the premises on which it depends. But analogical reasoning and anecdotal evidence do have a strong history in scientific inquiry, often as a catalyst to more experimental approaches to addressing the question at hand.

Both arguments from analogy as well as experimental data from animal behavior and cognition research are brought to bear on answering the question of whether some nonhuman animals possess mental states similar to the ones we take to explain human behavior. Analogical arguments are not deductively valid, but analogy does provide inductive support for inference. Although analogical inference may be justified when we have reason to suppose isomorphism (i.e. equality) between the causal structures of the two systems responsible for generating the features we are interested in, the argument from experimental data is needed to justify the initial similarity on which the analogical inference is based.

Although the problem of other minds is as much a problem for ascribing mental states to other humans as to nonhuman animals, such ascriptions are sanctioned in the case of the former but not the latter. Species membership grounds our belief that in the case of another human the relevant causal structures are not only isomorphic, but homomorphic (i.e the same, identical). If two structures are homomorphic, then there is just one type of causal structure being instantiated in each individual rather than two different types of causal structures that result in the same output. The ability to provide verbal reports of one's mental states as well as the shared structure of the brain and nervous system provide further justification for the inference of shared causal mechanism in the case of other humans. The objections for extending this inference of shared causal mechanism from human to nonhuman animals emphasize the fewer types of supporting evidence in the cross-species case: e.g. no verbal reports or shared species membership.

Those arguing against nonhuman animal minds question not only the similarities on which an analogy to the role human minds play in guiding behavior are based, but also the analogy itself, often arguing that the animals are not actually engaging in the behavior they are said to be engaged in. For example, when humans and nonhuman animals engage in behavior that indicates that they recognize the role that beliefs and desires play in influencing a conspecific's behavior, humans are taken to be reading the mind of that conspecific whereas other animals are presumed to simply be reading the conspecific's behavior.

III. Evidentiary standards: simplicity considerations

In the name of 'simplicity,' the principle of parsimony tells us that *entities should not be multiplied beyond necessity*, i.e. a hypothesis should not be asserted nor an entity postulated if it is not needed for explanatory purposes. But the principle of parsimony may be formulated in a variety of ways. Sometimes the requirement for simplicity in explanation is invoked in reference to the complexity of different kinds of entities being postulated (qualitative parsimony) and other times in regards to the actual number of those entities being invoked (quantitative parsimony).[1]

[1] In animal behavior and cognition research, quantitative and qualitative parsimony are often at odds. They advise the opposite entities to be dispensed with, i.e. a differentiating entity in cross-species comparisons (quantitative parsimony) or a common entity/psychological mechanism in explaining the behavioral repertoire of a species (qualitative parsimony). In animal behavior and cognition research, those providing associative accounts of apparently complex behavior must postulate numerous individual learning histories for each task the animal subjects perform, and those learning histories generally only account for the one experimental result they are discussing. This is in contrast to those proposing that the animal subjects' possession of higher-order cognitive processes better account for the findings, who claim that postulating those higher-order mechanism better explains the entire stock of experimental findings.

William of Ockham (1285-1347/9) is credited with codifying the *law of parsimony* (a.k.a. the "law of economy") in the 14th century and his 'Razor' is still taken as the standard justification for the epistemic virtue of simplicity in scientific explanation: *"pluralitas non est ponenda sine necessitate"* ("plurality should not be posited without necessity"). In other words, of two competing theories, all else being equal, the simplest is to be preferred. But the stipulation of "without necessity" indicates that it is not just the simplest of all explanations that could be posited, but the simplest explanation "that accords with our state of knowledge about the object or event in question" (Welsh, 1798).

As formulated by Ockham, the principle of parsimony counsels agnosticism; "[I]t tells us to remove what is unnecessary" (Sober, 1981, 145). But Elliot Sober contends that when we look at what scientists actually achieve when they apply the principle of parsimony to their hypotheses they do not simply remain agnostic about the existence of superfluous entities (i.e. reduce/assimilate), but they actually reject the existence of those entities (i.e. razor/eliminate). In this atheistic formulation, "The principle of parsimony counsels that *we should hypothesize that an entity does not exist*" (Sober, 1981, 145). According to Sober, although reduction also suggests that reducible posits should be dispensed with, reduction and razoring remain two different kinds of reducibility arguments. When we reduce we do not deny the existence of an entity, we just postulate that it is identical with another entity whose existence we accept. But when we razor, we ultimately dispense with the entity under consideration.

Sober proposes he can reunite parsimony and reduction; "If two things are *identical*, then *there is no property* which one has and the other lacks. An identity claim is a claim of nonexistence. Thus, reduction itself is a razoring of sorts" (149). That is, when we reduce we reject existence of a differentiating property between two entities and thereby eliminate that property from our explanatory scheme. Sober proposes that it follows from this that a rationale for reduction will follow from the justification of the principle of parsimony. If we can identify two entities with one another, it is reasonable to assume that they are identical; if we can dispense with a theoretical posit, then it is reasonable to think that posit does not exist. But Sober contends that,

> [T]he fact that an existence claim fails to increase the explanatory power of one's theory of the world is not sufficient for one to be reasonable in thinking it is false. There can be reasons for including a hypothesis in one's body of beliefs other than increasing that system's explanatory power. (150)

The agnostic formulation, which advises us to remove what is unnecessary, can be justified via a simple probability argument according to which the conjunction of two postulates is less probable than either of the postulates alone, but justification for the atheistic formulation of the razor, which advises the rejection of a differentiating property/entity, has not been provided.

> [T]he razor demands that a postulate be rejected if it, unlike its competitors, is not needed to explain *anything*; its superfluity in a particular case is not what matters. In the case that two existence claims each are able to explain a given phenomenon, we should reject the one that is not needed to explain any other phenomenon. The razor

> is thus nothing more than a principle of induction which focuses on existence claims. (Sober, 151)[2]

In regards to the problem of other minds, the agnostic version of the principle of parsimony, which sees reduction as assimilation, assumes isomorphism of causal structures underlying a given behavior (i.e. they are equal for present purposes, even if not identical). That is, it advises preference of the simplest explanation when the model and target system are identical for the present purposes. In contrast, the atheistic version of the principle of parsimony, which views razoring as elimination, requires evidence of homomorphism of causal structures (i.e. they are the same thing, and just one thing) for an entity to be dispensed with. But in examining the cognitive mechanisms underlying shared behaviors of humans and nonhuman animals, we must also allow for possibility of homoplasy (e.g. bird and bat wings are homplasies), not only homology (e.g. vertebrate limbs are homologies) when making claims of simplicity in explanation.[3]

---

[2] The second sentence of this quotation from Sober is particularly relevant to the debate over nonhuman animal minds because, as we will see in the following chapter, whereas as the naysayers advocate for their position on the basis of qualitative parsimony (i.e. the explain the findings by postulating numerous simpler cognitive mechanisms to account for each of the experimental findings), supporters of the attribution of mind to some nonhuman animals defend their position by demonstrating that it is quantitatively simpler to postulate a single complex cognitive mechanism (i.e. attributing mental representations to the animal in question) that explains the entirety of experimental data.

[3] Whereas homologous traits amongst two or more species are shared because they evolved in a common ancestor, homoplasious traits are the result of convergent evolution leading to similar traits amongst species that were not derived from a common ancestor. Convergent evolution is the independent evolution of analogous traits in species whose last common ancestor did not possess that trait.

IV. Methodological principles: Morgan's Canon

In *The Descent of Man* (1871), Charles Darwin placed animals and humans on a continuum, but the evidence he provided for this presumed similarity was predominately based on anecdotes collected from his peers. George Romanes, an assistant to Darwin in his later life, coined the term 'comparative psychology.' Romanes' *Animal Intelligence* (1882) also provided predominately anecdotal evidence in defense of his claims for the attribution of mental states to explain the complex behavior of some nonhuman animals.

In his *Introduction to Comparative Cognition* (1894), British psychologist C.L. Morgan presented a reaction to the anecdotal strategy utilized by Darwin and Romanes. Morgan's text defended the *double-inductive method*; Morgan drew a distinction between (1) objectively testable inferences from animal behavior and (2) untestable speculations about animal minds, christening the former as 'scientific' and the latter as 'unscientific.' According to Morgan, questions about the cognitive processes of nonhuman animals "will have to be settled, if... [they] can be settled at all, not by any number of anecdotes, -- interesting, and to some extent valuable, as such anecdotes are, -- but by carefully conducted experimental observations, carried out as far as possible under nicely controlled conditions" (1903, 359). Morgan worried not only about the dangers of anthropomorphism, but also was concerned with the human tendency to assume that our own behavior is underwritten by higher processes of reasoning when in many cases it too can be explained by simpler

mechanisms. He believed that deep and sustained introspection on the part of the psychologist could correct both of these tendencies.

Morgan's Canon, as first formulated in his1894 textbook, states; "*In no case may we interpret an action as the outcome of a higher psychical faculty if it can be interpreted as the outcome of the exercise of one that stands lower in the psychological scale*" (53). Morgan justified his Canon on the basis that evolutionarily ancient, species-general cognitive mechanisms are widespread in the animal kingdom whereas specialized, advanced cognitive skills are only found in a restricted number of species. He proposed that we should therefore first seek to explain any behavior in terms of the former and only resort to the latter if more basic mechanisms are unable to account for our observations.

According to Morgan, if an organism's behavior is the result of trial-and-error learning giving rise to a simple stimulus-response connection in the animal's nervous system, then no mind was needed for it to be carried out. It is important to keep in mind that Morgan did not reject the existence of higher order cognitive mechanisms in nonhuman animals, but that he did believe that the tendency towards anthropomorphism led to many explanations of observed animal behavior in terms of advanced cognitive processes that could just as well be explained by simpler mechanisms.[4] And he advocated for the use of careful and sustained introspection

[4] Morgan had a fox terrier named Tony who could open the gate of a garden he was placed in and Morgan recognized that to the observer it would appear that the dog acted on a kind of insight into the problem that faced him. But Morgan had observed Tony on a previous occasion moving along the picket fence, sticking his muzzle between the spaces and that he had by chance arrived at the picket where the latch was located. It was observations like this that led Morgan to reject the purely anecdotal strategy of his predecessors in the study of the mental evolution of nonhuman animals

into our own cognitive processes to determine if postulating higher processes is actually justified when evaluating the mechanisms underlying nonhuman animals' behavior.

Morgan identified a number of objections to his Canon that could be raised. Firstly, "that it is ungenerous to the animal" (Morgan, 53). He responds by stating,

> [T]his objection starts by assuming the very point to be proved. The scientific problem is to ascertain the limits of animal psychology. To assume that a given action may be the outcome of the exercise of either a higher or lower faculty, and that it is more generous to adopt the former alternative, is to assume the existence of the higher faculty, which has to be proved. (Morgan, 53)

We can see from this statement that he is not restricting the possession of higher cognitive processes to the human species, but is rather shifting the burden of proof for those making such attributions of higher mechanisms in their explanations of nonhuman animal behavior away from where it had been placed by those arguing from purely anecdotal evidence.

The second objection Morgan raises illuminates his understanding of the Canon as distinct from, and in opposition to, the principle of parsimony. He states,

> [A] second objection is, that by adopting the principle in question we may be shutting our eyes to the simplest explanation of the phenomena. Is it not simpler to explain the higher activities of animals as the direct outcome of reason or intellectual thought, than to explain them as the complex results of mere intelligence or practical sense-experience? (Morgan, 54)

---

because a single observation of an apparently complex behavior was not able to provide the observer with knowledge of the ways in which that behavior could have been shaped by prior trial-and-error learning, rather than by insight to the problem at hand.

And he responds, "But surely the simplicity of an explanation is no necessary criterion of its truth" (54) … "we are forced as men, to gauge the psychical level of the animal in terms of the only mind of which we have first-hand knowledge, namely the human mind" (55). It is evident from these statements that Morgan viewed the anthropomorphic explanations of nonhuman animal behaviors as the most parsimonious. Whereas parsimony in scientific explanation is concerned with the number of assumptions that must be postulated in order for a hypothesis to explain an observed phenomenon, the Canon as a principle of animal behavior and cognition research is concerned with the complexity of the psychological mechanisms being invoked to explain an observed behavior.

Despite Morgan's intention, and despite his avowed recognition of the principle of parsimony and his Canon being at odds with one another, he quickly recognized that many readers were conflating the two principles.[5] In 1903 he released a revised edition of the *Introduction to Comparative Cognition*, and in it the phrasing of the Canon was modified to state:

> In no case is an animal activity to be interpreted in terms of higher psychological processes, if it can be fairly interpreted in terms of processes which stand lower in the scale of psychological evolution and development. (59)

Not only did he eliminate reference to mental "faculties" in this revision in an attempt to distance himself from the faculty psychologists of the time, he also makes explicit reference to the process of "psychological evolution," expressing his belief in not

[5] Roger K. Thomas (1998, 2001) has chronicled the mischaracterization of the Canon as a principle of parsimony as well as the attempts made by some scholars to correct that misrepresentation from Morgan's initial introduction of it to the present.

only the continuity of physical characteristics across closely related species, but of the continuity of mental processes as well. In a further attempt to avoid misinterpretations of his position, Morgan added a caveat to the statement of the Canon, stipulating that the simplest explanation must accord with the greater body of knowledge on the subject matter. He states,

> To this, however, it should be added, lest the range of the principle be misunderstood, that the Canon by no means excludes the interpretation of a particular activity in terms of the higher processes, if we already have independent evidence of the occurrence of these higher order processes in the animal under observation. (Morgan, 1903, 59)

Together with his acknowledgment of the psychological continuity of related species, it follows from this caveat that we should not only take into account what we know about the expressed cognitive capacities of the species under investigation, but also what we know about phylogenetic relationships between species as well as what we know about the mental processes of those closely related species.

For Morgan, the Canon was meant to counteract anthropomorphizing tendencies because simplicity considerations alone would lead us to infer similarity between the mental states of humans and those of other animals. Morgan himself viewed anthropomorphic explanations of nonhuman animal behavior as the most intuitively plausible (parsimonious in this reductive sense) and presented his Canon as a guard against this bias to assume similarity. Although Morgan formulated the Canon as a caution against the Razor in terms of attributing mental states to other animals, most often the two methodological values are conflated. If the Canon was

intended as a guard against Ockham's Razor, then the Canon cannot be grounded on the basis of it's being a particular instance of the Razor.

V. Experimental approaches to studying animal minds

E.L. Thorndike's (1911) *Animal Intelligence* brought the study of animal minds (or the lack thereof) into the laboratory. Following Morgan, Thorndike concluded that apparently intelligent behavior could be carried out by (a series of) simpler associative mechanisms and that the inference to reason or consciousness was both unnecessary and misleading. It appears that Thorndike, not Morgan, equated the Canon to the Razor, proposing that postulating simpler mechanisms would result in the most parsimonious account of the data. This misreading of the relation between the general principle of parsimony and Morgan's Canon predominates to do this day.

In "Psychology as the Behaviorist Views It" (1913) John B. Watson proposed that there had been much progress in understanding simple associations; many experiments on conditioning generated a number of complex theories, but little or no reference to intervening mental processes was made. In *Behavior: An introduction to comparative psychology* (1914), Watson makes no reference to the Canon, but does mention Morgan's introspective method when he states, "Examination of the literature shows that experimenters have usually chosen some anthropomorphic type of classification of imitation, such as that outlined by Morgan" (278). Watson was reacting against the proposal by Morgan that the introspective processes of the psychologist studying nonhuman animal behavior constituted a scientific method.

Following Watson, Ivan P. Pavlov, in his *Lectures on Conditional Reflexes* (1928), was unwilling to propose unseen mental processes, but he did pose unseen physiological processes that could explain his observations. B.F. Skinner's radical behaviorism exemplified in *The Behavior of Organisms* (1932) provides an explicit dismissal of the notion that mental processes control behavior. He claims that all behavior, including mental images, can be explained solely by reference to environmental contingencies impinging on the human or animal. Skinner proposed a strict associationism in terms of cognitive processes, even for humans.

But, the predominately behaviorist orientation of experimental research on nonhuman animal behavior and cognition before 1960 is not the whole story. Not everyone rejected the existence of mental processes in some nonhuman animals. In *The Mentality of Apes* (1917), Wolfgang Köhler discussed insightful chimpanzees. Edward Tolman wrote about "Cognitive Maps in Rats and Men" (1948). In *A Textbook of Psychology* (1958) Donald O. Hebb argued that 'mind' is a name for processes in the head that control complex behavior, and that it is both necessary and possible to infer those processes from behavior. Ulrich Neisser's *Cognitive Psychology* (1967) exemplifies the beginning of the 'Cognitive Revolution' in research on humans and spurred similar transformation in research on animals.

VI. Ape language projects

Whereas comparative cognition researchers were predominately concerned with species-typical traits and making comparisons across species, those working in

the emerging field of primatology were more likely to recognize the importance of individual differences in their subjects. A group of these early primate researchers took a different approach to correcting for the indirect access to the internal mental states of nonhuman, nonlinguistic animals by attempting to teach their ape subjects human-like language. In the 1920s Robert Yerkes attempted to teach chimpanzee subjects to speak English.[6] His attempts were largely unsuccessful and he attributed this to their failure to imitate sounds. Others attributed his lack of success to physiological factors, specifically because chimpanzees' vocal apparatus is not designed for speech and that it is not under voluntary control.

In the 1960s another wave of ape language projects attempted to accommodate these physiological disadvantages by teaching ape subjects artificial languages that did not require them to speak. In 1966 Allen and Beatrice Gardner began teaching an infant chimpanzee named Washoe to use American Sign Language (ASL), relying on both imitation and instrumental conditioning. Washoe was able to transfer signs she'd learned directly from her trainers to novel items (for example, using 'more' in a variety of contexts not just for more tickling, which is the context she had been taught the sign) (Gardner and Gardner, 1989, 190). She also used the sign 'dog' to respond to the barking of an unseen dog (191). Additionally, the Gardners reported that after acquiring ten signs, she began combining them spontaneously. And when Washoe adopted an infant chimpanzee named Loulis, he

[6] After observing a colony of chimpanzees in Cuba in the early 1920s, Yerkes purchased two chimpanzees from a zoo. His book, *Almost Human* (1925), tells the story of the summer he spent with two chimpanzees, Chim and Panzee, who lived in a bedroom of his home.

managed to acquire over 50 signs even though the trainers had refrained from signing in his presence.

In 1967 David Premack started working with his chimpanzee subject, Sarah, on a token-based language system. Tokens of various colors, shapes, sizes, and textures represented nouns, verbs, adjectives, pronouns, and quantifiers. To test her understanding of the relationship between the tokens and the things signified, Premack presented her with an apple and a set of features (e.g. round vs. square; red vs. green) and then with her token for 'apple,' which happened to be a light-blue plastic triangle, and the same set of features. For both the object and the token, Sarah selected the features that describe an actual apple. (Anne Premack, 1976, 104).

The ape language programs have not been restricted to chimpanzees. Duane Rumbaugh and Sue Savage-Rumbaugh's bonobo subject, Kanzi, learned to use an electronic keyboard composed of lexigrams by watching his mother's lessons and was able to follow structured rules in multi-word sentences. Francine Patterson began teaching Koko the gorilla sign language in 1972 and their work together continues to this day. Koko has been observed to use signs to tell jokes and to communicate her inner feelings as well as to sign about things that are not immediately present.

All of these apes' trainers claimed that their subjects had learned hundreds of words and were able to string them together into meaningful sentences, as well as to have coined new phrases. But there were many reasons to doubt the extent of success claimed. According to Steven Pinker (*The Language Instinct*, 1994), most of these

trainers have made little to none of their raw data available to the research community. And when Herbert Terrence and colleagues tried to teach ASL to a relative of Washoe's named Nim Chimpsky, in addition to scrutinizing the available published data on the other signing apes, it was revealed that the apes had not really learned anything like ASL.

In terms of the number of words claimed to have been acquired, the trainers would translate pointing as a sign for 'you,' hugging as a sign for 'hug,' same with picking, tickling, kissing. In addition, the same movement was often credited with multiple meanings depending on context. And the chimpanzees' abilities for grammar were largely nonexistent. Inflection, which is the primary means of conveying who did what to whom as many other kinds of information in ASL, was absent from the chimpanzees' use of signs. (Pinker, 347)

Most of these psychologists have now abandoned their claims that their chimpanzee subjects acquired anything like a human language, other than Sue Savage-Rumbaugh and Duane Rumbaugh. The Rumbaughs have conceded that the chimpanzees did not learn much, but claim that bonobos do better. They claim that their research subject, Kanzi, has performed substantially better at learning graphic symbols and understanding spoken language than the common chimpanzees they worked with previously. But Kanzi spent his infancy observing his mother being trained on these same things, and her performance indicates that such training was largely unsuccessful. (Pinker, 350) David Premack moved on to an alternative research program aimed at probing the cognitive capacities of his chimpanzee subject,

Sarah. This research program investigating chimpanzees' ability to infer the mental states of others does not rely on linguistic communication between subject and researcher, but rather on the subject's observable behavior. It will be discussed in detail in the following chapter.

VII. Outline of project

A longstanding debate in the nonhuman animal cognition literature exists between those providing 'associative' accounts and those providing more 'mentalistic' explanations of nonhuman animal behavior and cognition. Whereas the former maintain that all nonhuman animal behavior can be explained as the result of conditioned associations or innate reflexes, the latter propose that some nonhuman animals possess the capacity for higher levels of mental representation in addition to the capacity for associative learning. The association/cognition debate, as it is understood in this project, represents the current incarnation of the animal minds debate. It emerged as such following what has been termed the 'Cognitive Revolution' in the human psychological sciences and the failure of the ape language projects. Following the hegemony of Skinnerian strictly behavioristic accounts of apparently mentalistic activities, which predominated from the 1950s until the 1980s, the question of whether humans where exceptional in their exemption from behavioristic accounts once again gained attention.

The current state of debate within philosophical circles if noteworthy for anything is for the lack of consensus on just about anything, including what the

critical questions to be answered are and what evidence would shed light on such questions. But interactions between divisions of the university are finally bringing theoreticians together with researchers and, at least, some new and interesting insights as to where exactly the debate centers (what is at stake; what is being advanced/contested by the various interlocutors in the debate) are being revealed. One area of inquiry that received significant attention in the latter part of the 20th century centers around the question of whether humans are unique in their ability to form representations of another's mental states, that is, whether all other animals rely on strictly associationist mechanisms that enable a kind of "behavior-reading" rather than "mind-reading." This debate has been posed in terms of a question of uncertainty, i.e. the subjects either do or do not possess a theory of mind, and the experimental apparatuses utilized in answering this question are intended to reveal what traits are or are not present in the species under investigation.

In Chapter 2 I will review this particular instantiation of the association/cognition debate and attempt to disentangle the causes of the dispute, often assumed by philosophers to revolve around disparate conceptions of simplicity considerations. I will propose that the naturalistic tasks put to the subjects is itself what allows the formulation of seemingly contradictory conclusions to be drawn from the experimental results. By examining only species-typical behaviors and the cognitive capacities underlying them, researchers are confronted with the possibility that those behaviors are the result of trained associations acquired in the animal subject's individual learning history, or stimulus-bound relations acquired in the

evolutionary history of the species to which they belong, rather than resulting from the subject's ability to consciously reason about it's environment (including the other individuals that are part of it).

In the primate mindreading debate, as it has played out, "positive findings" are in principle incapable of providing evidence because learning history could explain results by virtue of the fact that the object of study, if it exists, is presumed to be a species-typical trait. The move towards more naturalistic experimental paradigms has unfortunately led to a further stagnation in the association/cognition debate. Utilizing naturalistic experimental paradigms give the associationist an in to make the claim that the skill has a prior reinforcement history, whether onto- or phylogenetic. Because of this, we should explore the epistemic virtue that non-naturalistic (i.e. ecologically-invalid) experimental paradigms could afford in resolving the dispute over whether advanced forms of cognition are limited to the human animal. We should overcome the restricted lens of inquiry that looks only to demonstrations of presumed species-typical traits of nonhuman animals that are proposed by the cognitivists to depend on some form of metacognition.

Chapter 3 will explore the modularity metaphor of human cognition and how it has restricted our investigations into the cognitive capacities of nonhuman animals by assuming that the what is at stake in such investigations is the demonstration of species-typical traits. Evolutionary psychologists investigating the cognitive mechanisms underlying human rationality have proposed that most, if not all, reasoning and decision-making in their human subjects is carried out by domain-

specific mental modules. Richard Samuels and Stephen Stich have objected to this hypothesis of 'massive modularity' and have proposed instead that reasoning in humans is underwritten by two distinct neuronal systems, one of which is constituted by a number of domain-specific mental modules and another that is constituted by a set of domain-general problem-solving skills. I will argue that Samuels and Stich's 'middle-way' is better able to account not only for the body of experimental data coming from studies of rationality in human subjects, but also for the results from studies of marine mammal cognition. Ronald Schusterman and colleagues' demonstration of the eventual acquisition of an equivalence concept by their sea lion subject, Rio, in a nonnaturalistic experimental paradigm will be provided as a case-study for the attribution of a dual-processing cognitive system to a nonhuman animal.

Positive results in an ecologically invalid experimental paradigm, like the one carried out on Rio, show that we do have experimental evidence of abstract thinking in a nonlinguistic, nonhuman animal. It has provided a sign of other minds not in verbal behavior/language, but in something that is indisputably rational (and thereby internal) because the animal subject lacks both the phylogenetic and ontogenetic history for such an ability to be the result of either operant conditioning or an innate releasing mechanism. I will propose that animal behavior and cognition research could benefit from a shift of focus from the demonstration of species-typical cognitive traits to an investigation of the potentials of nonhuman animal minds. By investigating the permeability of domains of reasoning rather than simply attempting to demonstrate species-typical traits (resulting from evolutionary psychology's

attempt to delineate the limits of their presumed domains), animal behavior and cognition researchers may illuminate important aspects of the structure of both human and nonhuman animals' minds.

Chapter 4 will provide a detailed account of how the concept of association has been used in both philosophical and psychological approaches to the study of the mental processes of both humans and other animals. Broadly speaking, 'associationism' refers to psychological theories that attempt to explain apparently complex cognition as being built-up from simple associations between sensations/stimuli and behavior/responses. Eric Mandelbaum has provided a useful discussion of the varieties of associationist theses (and the conflation of them) that have appeared in the human social psychological literature. In addition to distinguishing a variety of types of associationist theses, Mandelbaum also provides a method for determining whether or not some behavior is in fact underwritten by associative mechanisms; he does this by seeing how purported associations change, or do not change, under certain conditions.

Association is purported to be able to account for learning, the structure of mental states, and the way certain thoughts relate to other thoughts. By positing a single mental process underlying thinking, learning, and cognitive structure, the associationist purports to have parsimony on their side (at least, in a certain sense of parsimony). But, as Mandelbaum points out, accepting an associative account of one of these mental processes does not entail that we must also accept associative accounts of the others. The primate theory of mind literature will be reexamined in

light of this more thorough understanding of the types of novel behaviors that can and cannot be explained on a purely associative account.

Rethinking what the Canon advises in regards to the entirety of the primate theory of mind findings, and given that we can now attribute abstract reasoning (and the possession of internal thought, on which it depends) to a nonhuman animal, the burden of proof should be lifted from those who seek to explain numerous diverse complex behaviors in a nonhuman animal species by reference to higher order cognitive mechanisms when doing so accounts for the larger body of data, rather than advising them to explain a single experimental result in isolation by postulating a complex web of conditioned associations for which no evidence is available to show that such associational learning has taken place.

Morgan's Canon advises that nonhuman animals' behaviors be explained in the simplest (phylogenetically widespread) psychological terms when there is an absence of independent evidence of higher-order mechanisms operating in the species under investigation. But looking only at what a species does, even if in its natural environment, tells us nothing about what an individual animal subject can do under altered circumstances and when given the appropriate training. It also has the unfortunate consequence of leaving open the possibility that the given behavior is the result of stimulus-bound associations acquired in the organism's learning history. So, despite requiring the postulation of numerous assumptions about the organism's learning history providing the opportunity for such conditioned responses, associational accounts may be available to account for the observed behavior.

Ecologically-invalid, nonnaturalistic experimental paradigms eliminate such assumptions from consideration and thereby provide the strongest evidence of higher-order cognitive mechanisms operating in a nonhuman animal subject, even if they do not provide evidence that that trait is typical of the species in its natural environment.

Ch. 2: Uncertainty of Animal Minds: "mindreading" animals?

I. Association/cognition debate

A debate in the nonhuman animal cognition literature exists between those providing 'associative' accounts, which reference only the role of associative conditioning, and those providing more 'mentalistic' explanations of nonhuman animal behavior and cognition. The latter propose that some nonhuman animals possess the capacity for higher levels of mental representation in addition to the capacity for associative learning, on which the former presume all nonhuman animal cognition depends. Both propose to be providing the most 'parsimonious' account of the experimental findings.

One area of inquiry that received significant attention in the latter part of the 20th century centers around the question of whether humans are unique in their ability to form representations of what another individual knows (or does not know), that is, whether all other animals rely on strictly associationist mechanisms that enable a kind of "behavior-reading" rather than "mind-reading." Whereas mind-reading (is presumed to) require the ability to form mental representations of another individual's mental states, behavior-reading requires only the ability to respond to strictly perceptual cues in anticipating another's behaviors.

Some theorists have proposed that the development of the capacity for mindreading depends on linguistic competency[7], but others have argued that it is the testing paradigm itself that requires linguistic competence. Those who propose that language development is a necessary prerequisite for the development of a mindreading capacity pay little attention to the fact that the nature of the experimental tasks, which often require the subject to respond verbally, could be responsible for the results which indicate that mindreading capacities emerge in human children around age four, after the "grammar explosion" that typically occurs in the third year of life. The experimental evidence reviewed here will reveal that some mindreading capacities have been discovered to be operating in some non-linguistic creatures. Therefore, language cannot be a necessary prerequisite for the development of some elements of the mindreading faculty.[8]

---

[7] See Karen Milligan, Janet Wilde Astington, and Lisa Ain Dack, "Language and Theory of Mind: Meta-Analysis of the Relation Between Language Ability and False-belief Understanding," *Child Development,* 2007, vol. 78, number 2, pp. 622-646. The authors performed a meta-analysis of 104 studies reporting on a total of 8,891 human subjects under the age of 7. They examined the relation between children's language ability and their false-belief understanding. The ability to attribute false beliefs to another individual is psychologically interesting because it involves an understanding that others may hold beliefs that are different than one's own, and act on them. Such actions cannot be explained by reference only to the attributor's own true beliefs about the situation in cases where those true beliefs would lead to an alternative behavior than the false belief would lead to. It is presumed that the ability to attribute a false belief to another individual requires the attributor to mentally represent the contents of that other individual's beliefs in order to predict the behavior that it gives rise to.

The authors found that language ability is more predictive of later success at false-belief tasks than is success at false-belief tasks in predicting later language ability. But it is worth noting that their analysis does not rule out the possibility that sophisticated language use may not be a prerequisite for understanding that others have mental states which may be different than their own, but that language ability is simply necessary for the subject to appropriately respond to the experimental task itself, which often require the subject to express their response in sophisticated language. They begin with a presumption of causation and so only examine the direction of fit between advancements in language and the emergence of theory of mind capacities.

[8] See Jill de Villiers, "The Interface of Language and Theory of Mind," *Lingua*, 2007, 117 (11): 1858-1878. de Villiers has proposed an interrelation between language ability and mindreading. On de Villiers' account, initial word learning depends on conceptual developments that are supported by

In this chapter I will review this particular instantiation of the association/cognition debate and attempt to disentangle the causes of the dispute, often assumed by philosophers to revolve around disparate conceptions of simplicity considerations. For example, Simon Fitzpatrick has provided compelling reason to see the stagnation of the animal minds debate as resulting from the disputants adhering to two different kinds of simplicity considerations in evaluating and interpreting the experimental results: simplicity as psychological unity across species (quantitative parsimony) and simplicity as parsimony of mental representations (qualitative parsimony).[9] Elliot Sober has referred to these disparate conceptions of simplicity in terms of the distinction between the presumed virtue of postulating a unified model, which applies the same explanation to multiple data sets, and a disunified model, which applies a different explanation to each data set.[10]

I will propose that the naturalistic tasks put to the subjects is itself what allows the formulation of seemingly contradictory conclusions to be drawn from the experimental results. By examining only species-typical behaviors and the cognitive capacities underlying them, researchers are confronted with the possibility that those behaviors are the result of trained associations acquired in the animal subject's

---

proto-mindreading capabilities, but refinement of those latter capabilities is supported by the development of the language faculty, specifically the acquisition of propositional attitude terms. That is, language development assists higher-order reasoning about others' mental states. de Villiers admits that no experimental evidence is available to determine the direction of influence between language and theory of mind in the developmental period from age two to four in human children, proposing that more work with nonverbal tasks as well as with special populations (i.e. those with impairments of either or both of these capacities) may make available a more precise account of this interrelation.

[9] Fitzpatrick, Simon. "The primate mindreading controversy: a case study in simplicity and methodology in animal psychology," In *The Philosophy of Animal Minds*. Cambridge: Cambridge University Press, 2009; 258-277.

[10] Elliot Sober, "Parsimony and models of animal minds." In *The Philosophy of Animal Minds.* Cambridge: Cambridge University Press, 2009: 237-257.

individual learning history, or stimulus-bound relations acquired in the evolutionary history of the species to which they belong, rather than resulting from the subject's ability to consciously reason about it's environment (including the other individuals that are part of it). I will also offer suggestions for future research protocols and experimental methods aimed at addressing the actual cause of the stagnation of the debate and will advocate the use of non-naturalistic ecologically-invalid experimental paradigms to probe nonhuman animals' meta-cognitive abilities.[11]

II. Investigating primate minds

After the dominance of behaviorist explanations of the mid-20th century, the idea that there could be internal thought in the absence of external speech reappeared as a worthwhile experimental question with the work of primatologists David Premack and Guy Woodruff (1978).[12] Premack and Woodruff performed a series of studies on an adult female chimpanzee, Sarah, in which they probed her ability for problem-comprehension. This work developed out of as well as built on the investigations of chimpanzees' problem-solving skills pioneered by Gestalt psychologist, Wolfgang Köhler. Köhler was the first psychologist to conduct formal experiments (rather than simple observation) with the aim of determining whether chimpanzees engage in intelligent behavior. As stated by Köhler,

[11] Ecological-validity in experimental design refers to the degree to which the experimental setting and the nature of the task put to the subject mimic that organism's ecology (i.e. natural habitat and typical behaviors). Ecological-invalidity, in contrast, refers to the degree to which the experimental paradigm departs from the organism's normal way of life.

[12] David Premack and Guy Wilson, "Does the chimpanzee have a theory of mind?" *Behavioral and Brain Sciences* 4 (1978): 515-526.

> [W]e do not speak of behavior as being intelligent, when human beings or animals attain their objective by a direct unquestionable route which clearly arises naturally out of their organization. But we tend to speak of 'intelligence' when, circumstances having blocked the obvious course, the human being or animal takes a roundabout path, so meeting the situation. (Köhler, 1925, 3-4)

In *The Mentality of Apes* (1925), Köhler presented his findings on chimpanzees' ability to learn in the absence of explicit training, proposing that the results of his experiments are best explained by attributing to the chimpanzee subjects a moment of understanding and appreciation of the whole of the situation. He was interested in whether or not chimpanzees utilize insight in order to solve a problem; he sought to answer the question of whether intelligent behavior exists among nonhuman apes. His experiments examined the chimpanzee subjects' use and construction of tools in problem-solving tasks, particularly in the context of obtaining out of reach food.[13] He

[13] While marooned at a primate research facility maintained by the Prussian Academy of Sciences in the Canary Islands from 1914-1915, due to the outbreak of the First World War, Köhler had the opportunity to conduct a series of experiments on nine captive chimpanzees of various ages. The experiments were performed in a large outdoor pen that contained a variety of objects including boxes, poles, and sticks. Köhler constructed numerous problems for the chimpanzees to solve that involved obtaining food that was not directly accessible to them. The simplest task involved the experimenter placing food on the opposite side of a barrier from the subject. In these tests the chimpanzee subject observed the experimenter toss the food out of a window in the barrier, after which the window was shut and the food was visually inaccessible to the subject. The chimpanzees immediately moved away from the goal (i.e. the food) to circumvent the barrier and retrieve it, taking the shortest possible indirect route to obtain their goal.
Other problem tasks put to the chimpanzees involved bananas hanging out of reach overhead of the chimpanzees. Although the details of the chimpanzee subjects' solutions varied, a typical sequence emerged: after a period of unsuccessful jumping, the chimpanzee subject appeared to become angry or frustrated, walked away from the problem task, paused, and then looked back at the bananas in a seemingly reflective way, then looked towards the objects in the enclosure (boxes, sticks, and poles), back at the bananas, and then looked back at the toys, after which the chimpanzee subject would proceed to use the objects available in an attempt to retrieve the bananas. Some of the chimpanzees stacked boxes underneath the bananas and attempted to climb up to the bananas, one tried unsuccessfully to shimmy up a pole he had placed under the bananas, and another moved a single box under the bananas and then used a pole to knock them down. To Köhler it appeared evident that the chimpanzees were experimenting in their minds before manipulating the tools, i.e. that they were solving the problem via a kind of mentalistic trial-and-error before performing the sequence of actions.

found that his chimpanzee subjects could assemble two sticks into a longer instrument in order to reach bananas that were out of reach overhead. His chimpanzee subjects also demonstrated an ability to pile crates on top of one another in order to climb up to obtain the bananas, and even combined these techniques when necessary. All of this was done spontaneously and in the absence of explicit training.

Premack and Woodruff expanded the question of whether chimpanzees behave intelligently to include the question of whether chimpanzees understand that other individuals have mental states (i.e. thoughts, beliefs, desires, …) that may be different than their own. In order to demonstrate that an animal subject is able to represent the mental states of another individual experimentally, one must determine whether they can represent mental states that are different than their own. By looking at only shared mental states, it is not possible to rule out the possibility of an absence of mental states altogether and that an apparent ability to attribute thoughts, beliefs, desires, … to another individual is being carried out non-representationally. They sought to find evidence of nonhuman animal minds by demonstrating that an organism is capable of reading the mind of another creature, specifically, of a human experimenter. In an attempt to determine whether their chimpanzee subject Sarah was able to comprehend a problem faced by a human experimenter, Premack and Woodruff had her observe a human actor confronting inaccessible objects and then asked her to indicate how the human actor would solve the problem.

He concluded that his chimpanzee subjects were utilizing insight in arriving at a solution to the problem at hand.

In order to test whether their subject comprehended the problem faced by the actor they created four short videotapes of a human actor in a cage struggling to obtain bananas that were (1) attached to the ceiling, so out of reach overhead, (2) outside the cage wall, so horizontally out of reach, (3) outside the cage, but the actor's reach was impeded by a box inside the cage, and (4) outside the cage with actor's reach impeded by a box, and the box was filled with heavy cement blocks (Premack and Woodruff, 516). They also produced still photographs of the actor engaged in an appropriate solution to each of the problems shown in the videos: photographs of the actor (1) stepping onto a box, (2) lying on his side and reaching out of the cage with a rod, (3) moving a box to the side, (4) removing cement blocks from a box (Premack and Woodruff, 516). During testing, Sarah was shown one of the videos and would then be provided with a choice of two still photographs, one of which depicted the actor engaged in the behavior that constituted a solution to the problem faced in the video she had just viewed. She was trained to identify her selection by placing it in a designated location beside the television set on which she viewed the videos. 6 trials were performed on each of the 4 problem-solution tasks. Sarah was correct on 21 of 24 trials and her errors were confined to the problem that required the actor to remove the cement blocks from the box blocking their access to the bananas before attempting to move it out of the way. Of the six trials on this problem, Sarah made three errors in a row but then succeeded on the last three trials by selecting the photograph of the actor removing the cement blocks from the box. Sarah selected the correct response for all 6 of the trials on each of the other problems. Whereas Köhler

had presented the problems to his chimpanzees to solve themselves rather than to identify the behavior a human actor would need to engage in to solve the problem, the heavy box was the problem with which Köhler's chimpanzees had the greatest difficulty as well.

Premack and Woodruff presented three possible interpretations of their chimpanzee's comprehension of the problem faced by the human experimenter: associationism, theory of mind, and empathy. On the associationist interpretation, the subject selected the correct solution on the basis of her familiarity with the sequence of behaviors carried out by the actor in the videotape and then selecting the photograph of the image that completed that sequence. On this reading of the results, it is the similarity between past cases and the present one that allow the animal subject to select the appropriate response (Premack and Woodruff, 518). The researchers admit that they are unable to rule out the associative interpretation of the results because, although the chimpanzee subject had never received explicit training in how to solve the given problems, it is possible that she had seen a human engage in such behavior at some point in her past. Nevertheless, they present the interpretation that attributes mind-reading abilities to their subject as the most compelling of the three. On this account, the chimpanzee selects the appropriate solution to the problem faced by the human actor by imputing at least two states of mind to that actor: intention (i.e. that the actor wants the banana and is struggling to reach it) and knowledge (i.e. that the actor knows how to attain the banana).

They take the final alternative, what has come to be known as simulationist theory of mind, but what they call the "empathy interpretation," as the least compelling of the three. [14] On the empathy account, the chimpanzee subject selects the appropriate solution to the problem by putting herself in the actor's place and selecting the solution that she would carry out if faced with that problem. Rather than making a prediction of what the actor would do on the basis of what she takes the actor to know about the situation, she is instead posited to be simulating what she herself would do in such a situation based solely on her own knowledge of the situation. As stated by Premack and Woodruff, "empathy and 'theory of mind' are not radically different views; they are in part identical. … The empathy view diverges only in that it does not grant the animal any inferences about another's knowledge; it is a theory of mind restricted to purpose" (Premack and Woodruff, 518). That is, on the empathy interpretation the chimpanzee is able to infer the actor's motivation, but not their knowledge and or beliefs about the situation they are confronted with. So,

[14] Two different models of theory of mind have been posited, the second as a response to the first. As originally conceived by Premack and Woodruff, the ability to predict and explain behavior of oneself and others is underpinned by a folk-psychological theory of the structure and function of the mind. This model is referred to as theory-theory. On the theory-theorists' account, individuals acquire and deploy a commonsense (i.e. folk-psychological) theory of mind, something akin to a scientific theory in that it postulates unobservable entities in the service of predicting and explaining observable phenomenon, in order to accomplish the mindreading process.
According to simulation-theory, at the root of mature mindreading lies an ability to project one's self into another individual's perspective. Mindreading, on this view, involves a process of simulating the mental activity of another individual. Simulationist accounts do not explain mindreading capacities as dependent on some sort of theoretical knowledge of how an agent's beliefs and desires inform their behavior. As stated by Gallese and Goldman, "Simulation-theory suggests that attributors use their own mental mechanisms to calculate and predict the mental processes of others" (496). The individual attributing mental states to another individual does so by putting themselves in that other individual's shoes. On this account, mindreading involves an individual's mentally replicating how they would feel and react if in that other individual's place with no need to posit the unobservable inner states of another individual's mind.

instead, she uses her own knowledge and beliefs about the problem to select the appropriate solution.

Premack and Woodruff were unable to rule out either of the alternative interpretations of their findings, i.e. association and empathy.[15] But their work introduced questions about the mental states of nonhuman animals that were amenable to experimental investigations, despite the absence of a shared language between researcher and subject. Before we can proceed, we need to become clear about how the possession of a theory of mind relates to the capacity for higher-order cognition.

III. Theory of Mind (ToM)

The term 'theory of mind' was introduced by Premack and Woodruff in their 1978 paper reporting the results of their problem-solving experiments with chimpanzee subject, Sarah. Theory of mind (henceforth ToM), also known as *folk psychology* or *mindreading*, refers to "a system of knowing that enables individuals to infer what others believe, desire, and want" (Hauser, 61). It involves an understanding of the fact that others possess mental states that direct their actions and behaviors. As explained by Peter Carruthers and Peter Smith, ToM is a research domain whose goal is to provide an explanation of the ability to explain and predict the actions both of one's self and of other intelligent agents (1).

[15] Premack and Woodruff admit that their interpretation of the findings is underdetermined by evidence because, unlike in the case of linguistic production, the data are vague and the behavioral sequences are ill defined. They propose that future research should aim to resolve these issues so as to strengthen the basis of the theory of mind interpretation.

As conceived by Premack and Woodruff, the deployment of a human folk psychology may be understood as theory-like. It involves the use of a set of inferences about things that are *not directly observable*, i.e.: intentions, desires, thoughts, and beliefs. As Premack and Woodruff put it,

> In saying that the individual has a theory of mind, we mean that the individual imputes mental states to himself and to others (either to conspecifics or to other species as well). A system of inferences of this kind is properly viewed as a theory, first, because such states are not directly observable, and second, because the system can be used to make predictions, specifically about the behavior of other organisms. (15)

On their account, ToM may be utilized by individuals to make predictions about the future behavior of both themselves and others. In their attempts to develop more decisive ways of testing their hypothesis, Premack and Woodruff introduced numerous questions as to the extent to which chimpanzees have a human-like ToM, but provided few answers.

IV. False-belief task

Currently, the primary method utilized to uncover whether an organism possesses elements of a ToM is the 'false-belief task.'[16] In a standard false-belief task experiment, a human child watches as a character puts an object in a location in the

[16] The false-belief task experimental paradigm was introduced by Wimmer and Perner and presented in their paper, "Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception," *Cognition*, 13, 1983, pp. 103-128. In the years since, numerous permutations of the false-belief task have been developed with the aim of ensuring that successful performance does not require advanced language skills. Whereas the standard false-belief task requires an elicited response by the subject (i.e. they must respond to a direct question about the agent's false belief), the later permutations of the task depend on a spontaneous response by the subject (i.e. looking time or gaze direction).

presence of a target subject. The target subject then exits the room. At this point the child witnesses the character moving the object to a different location. The child is then asked where the target subject will look for the item when they return. In order to answer the question correctly the child must be able to distinguish perceptions from reality; in this case, they must be able to differentiate between the subject's perceptions and their own knowledge of the object's location. It is generally agreed that success at a false-belief task reveals the possession of ToM, but the primary problem with this experimental paradigm is its reliance on verbal communication between the experimenter and the test subject. Because of this requirement, the standard false-belief task paradigm has been restricted to use with language using beings.

Kristine Onishi and Renee Baillargeon overcame this restriction of the false-belief task paradigm in the experiments they performed on pre-linguistic human infants.[17] To Onishi and Baillargeon, possession of ToM is required for an infant to understand that other's behaviors are based on their beliefs, and that those beliefs may be true or false. They describe ToM as a being's awareness that beliefs are representations, and that representations are not necessarily direct reflections of reality. In order to examine the infant's capacity for such knowledge they introduced a novel non-verbal task in order to determine if 15-month-old human infants have the ability to predict an actor's behavior on the basis of their true or false belief about a toy's location. Onishi and Baillargeon utilized a 'violation-of-expectation' model in

[17] Kristine H. Onishi and Renee Baillargeon, "Do 15-Month-Old Infants Understand False Beliefs?" *Science*, 308, 2005, pp. 255-258.

which the infant's looking time was used to determine if they were surprised by the actor's behavior. The violation-of-expectation experimental method involves showing a subject (usually a prelinguistic human infant) an expected event and an unexpected event. Increased attention to the unexpected event is taken as evidence that the subject is surprised by the deviation from their own expectations.

The infants watched as an actor hid a toy in one of two locations. Then, a change occurred which resulted in the actor holding either a true or a false belief about the toy's location. They state, "If the infants expected the actor to search for her toy on the basis of her belief about its location, rather than on the basis of (their own knowledge of) its actual location, then they should look reliably longer when that expectation was violated" (Onishi and Baillargeon, 256). In other words, the infants should look longer when the actor searches for the toy in the location that does not correspond with the actor's belief, whether that belief is true or false. Onishi and Baillargeon's subjects looked on average ten second longer when the actor's behavior and belief did not match (i.e. when the actor looked where the toy was actually hidden) than when the actor searched in the location that correlated with their false belief about the toy's location. The results of the experiment revealed that infants based their expectations of the actor's behavior on that actor's beliefs, not on their own knowledge of the location of the toy. These results were interpreted by Onishi and Baillargeon as evidence that 15-month-old human infants possess a ToM in that they appear to understand that false beliefs lead to actions that may be different than those that a true belief would lead to.

With their non-verbal false-belief task Onishi and Baillargeon were able to identify evidence for ToM in human infants far below the age that the standard false-belief task experiments were able to. By introducing an experimental paradigm that does not rely on verbal communication between the experimenter and the test subject, they introduced evidence for ToM in human infants that do not yet possess linguistic abilities. This opened the possibility not only for studying ToM in non-linguistic animals, but it also revealed that a developed language faculty is not a necessary prerequisite for the development of ToM in non-human animals.

Higher-order-thoughts are thoughts that have mental states as their objects. It is often presumed that the 'vehicle of thought' required for such a mental operation must be linguistically structured, but evidence coming from sophisticated developmental psychology studies of prelinguistic infant cognition performed by Onishi and Baillargeon indicates that it is not language use *per se* that is the necessary prerequisite. The possibility remains that a related skill that underlies the ability to acquire language also facilitates meta-cognitive processing. But meta-cognition does not appear to be dependent on the possession of a functioning language capacity.

V. Primate mindreading debate

In their 1997 text, *Primate Cognition*, Michael Tomasello and Josep Call reviewed the available evidence for ToM capacities in non-human primates.[18] They arrived at the conclusion that some primates display some understanding of visual

[18] Michael Tomasello and Josep Call, *Primate Cognition*, Oxford University Press, 1997, pp. 311-341.

perception and that there is evidence that they use bodily orientation when deciding when a communicative signal should be given. But they go on to assert that there is no solid evidence that any non-human primate has an understanding of intentionality or of the mental states of others. But in a 2003 article they recant this position.[19] Although they do not propose that non-human primates have a full-blown ToM, they do admit that they are now convinced that some non-human primates do understand some psychological states in others. They do not identify which psychological states these are, but instead argue that a goal of future research should be to identify what mental states non-human primates do understand as well as to what extent they understand them.

The breakthrough experiment that changed Tomasello and Call's minds involved pitting a subordinate and a dominant chimpanzee against one another in competition over food in a situation in which only the subordinate had information about the location of a second piece of food.[20] The subordinate chimpanzees took advantage of this situation in very flexible ways. The subordinate chimpanzees avoided the food that the dominant chimpanzee could see and instead pursued the food that she had not seen. Tomasello, Call, and Hare (2003) propose that these experiments revealed that chimpanzees know something about the contents of what others see, and in some situations they seem to understand how what another sees affects their behavior (155).

---

[19] Michael Tomasello, Josep Call, and Brian Hare, "Chimpanzees understand psychological states- the question is which ones and to what extent," *Trends in Cognitive Science*, vol. 7, no.4, 2003, pp. 153-156.

[20] B. Hare, J. Call, B. Agnetta, M. Tomasello, "Chimpanzees know what conspecifics can and cannot see," *Animal Behaviour*, vol. 59, 2000, pp. 771-785.

Call, Hare, Carpenter, and Tomasello (2004) performed another series of experiments, which addressed the question of whether chimpanzees understand intentional action.[21] In these experiments an experimenter holding food was either unwilling or unable to give the food to the chimpanzee. The chimpanzees were more impatient and seemingly angrier with the experimenter that was unwilling to give them the food than they were with the experimenter who was unable. According to Call et al. this study provides evidence that chimpanzees can discriminate between intentional and accidental actions. But these findings contrasted sharply with the findings of earlier studies conducted by Daniel Povinelli.

Until the year 2000 there had been a general consensus that all non-human animals, including chimpanzees, lacked all of the components of human folk psychology. This conclusion was supported by the experimental studies conducted by Povinelli and Eddy (1996). In Povinelli and Eddy's experiment, a chimpanzee entered a test room in which two human experimenters were present. One of the experimenters could see the chimpanzee and the other could not because of either their body position, a blindfold over their eyes, or a bucket placed over their head. The chimpanzees in these studies never learned to beg selectively from the experimenter who could see them, leading Povinelli to conclude that chimpanzees do not use knowledge about what another individual has visual access to in order to infer what that individual knows.

[21] J. Call, B. Hare, M. Carpenter, and M. Tomasello, "'Unwilling' versus 'unable': chimpanzees' understanding of human intentional action," *Developmental Sciences*, vol. 7, no. 4, 2004, pp. 488-498.

However, Povinelli's conclusions were challenged by Brian Hare and colleagues (J. Call, B. Agnetta, and M. Tomasello in 2000; J. Call and M. Tomasello in 2001). Rather than utilizing an experimental design necessitating cooperative communication between a chimpanzee and a human experimenter, Hare and colleagues designed an experiment in which two chimpanzees of different dominance ranks were involved in a competition over food. Hare et al.'s experiments were intended to test whether chimpanzees use information about what a conspecific sees to infer what they know. Hare found that if a subordinate chimpanzee recognizes that the dominant chimpanzee fails to see the baiting of the food, then that subordinate chimpanzee will retrieve that food. From this the experimenters concluded that chimpanzees possess some elements of our human folk psychology, but they are not unique among non-human animals in doing so.[22] [23]

[22] Recent experimental work indicates that gaze-following (the ability to follow the direction in which another individual is looking to a location in space) and its requisite elements of theory of mind are present in numerous domesticated animals. For example: in dogs (see Horowitz, "Theory of mind in dogs? Examining method and concept," *Learning and Behavior*, vol. 39, 2011, pp. 314-317), in goats (see Kaminski et al., "Domestic goats, *Capra hircus*, follow gaze direction and use social cues in an object choice task," *Animal Behaviour*, vol. 69, 2004, pp. 11-18), and in pigs (see Albiach-Serrano et al., "The effect of domestication and ontogeny in swine cognition (*Sus scrofa scrofa* and *S. s. domestica*)." *Applied Animal Behaviour Science*, vol. 141, 2012, pp. 25-35).

[23] Evidence coming from the field of experimental cognitive ethology indicates that hominoids (i.e. members of the ape superfamily, both extinct and extant) are not unique in either their capacity for second-order intentionality. Michelle Maginnity (2007) has provided some of the least contentious evidence to date for the attribution of a capacity for second-order intentionality and theory-of-mind (ability to attribute mental states to others) to a nonhuman animal. Maginnity utilized the Knower-Guesser paradigm first developed by Povinelli et al. (1990) for use with chimpanzees. Maginnity sought to determine whether dogs are able to understand that the visual perspective of a human informant has consequences for their knowledge of the situation. In the Knower-Guesser paradigm two human informants, one knowledgeable and one ignorant, indicate the location of hidden food to the subject. Maginnity conducted a series of four experiments. In the first, second, and third experiments, the dogs showed a significant preference for the informant that had been attentive during baiting. In the fourth experiment, which operated as a control, the dogs showed no preference between the informants when they had equal perceptual access to the baiting. Maginnity states, "Overall, the results across the four experiments provide strong evidence that the dogs responded on the basis of

VI. ToM in monkeys

In experiments that confirmed these results Jonathon Flombaum and Laurie Santos (2005) have examined the capacity of rhesus monkeys to deduce what others perceive based on where they are looking.[24] Following Hare's experimental design, Flombaum and Santos designed an ecologically relevant ToM task, in which a monkey could "steal" a grape from one of two experimenters, with which to examine the cognitive capacities of free ranging rhesus monkeys. As stated by Flombaum and Santos,

> Our work builds on the recent insight that primates will most likely exhibit sophisticated ToM abilities in experimental scenarios that mimic the natural situation for which these abilities have evolved-namely, competitive foraging situations. (447)

By presenting their subjects with an experimental task that they could be evolutionarily as well as developmentally equipped to solve, the researchers seek to identify the mental operations underlying a species typical trait in the context in which that trait would be utilized by the subject. In doing so, these researchers are providing a more charitable, i.e. ecologically-valid, experimental setting for their subjects.

Flombaum and Santos performed six experiments with the free ranging rhesus monkeys from the Cayo Santiago population in Puerto Rico. They sought to

---

what the informant had or had not seen during the baiting – consistent with the hypothesis that dogs have a functional theory-of-mind in their interactions with humans." (Maginnity, 121)

[24] Jonathon I. Flombaum and Laurie R. Santos, "Rhesus Monkeys Attribute Perceptions to Others," *Current Biology*, vol. 15, 2005, pp. 447-452.

determine whether these monkeys reason about what a human competitor can and cannot see. The test subjects were presented with the opportunity to "steal" a grape from one of two human "competitors." The researchers hypothesized that the subjects would be motivated to take the grape only if they could do so without being detected, and therefore they predicted that the subjects would use information about where the two competitors were looking when deciding which grape to approach. If the monkeys possess knowledge of the connection between perceptions and knowledge, then they should selectively approach the experimenter who was looking away from the "contested" grape or whose view of the grape was somehow blocked. The test subjects reliably retrieved the grape from the human competitor who had either their back turned to the grape, was not facing the grape directly, had their head and eyes oriented away from the grape, had only their eyes oriented away, or had a barrier blocking their view of the grape (Flombaum and Santos, 447). The rhesus monkeys tested reliably avoided the experimenter who could see the contested food item.

According to Flombaum and Santos, these experiments establish the first experimental evidence that a non-ape species may spontaneously reason about another individual's visual perception (449). They argue that the previous failure of non-human primates in studies of ToM capacities have been due to the use of cooperative rather than competitive experimental designs. As they state,

> The success of the experiments presented here is surely due in part to the fact that they mimic the socio-cognitive problems that primates naturally face in the wild. Specifically, they explore what primates

> know about the eyes of others through competitive foraging situations. (Flombaum and Santos, 450)

This species of monkey has knowledge not only of where others are looking, i.e.: gaze following, but they also are able to identify what others see.[25]

Laurie Santos, along with Aaron Nissen and Jonathon Ferrugia (2006), performed another series of experiments in which they sought to determine whether rhesus monkeys are aware of what others can and cannot hear. Specifically, they explored whether this species understands the connection between seeing and hearing.[26] The test subjects came from the same population as the subjects in the previous study investigating rhesus monkeys ability to infer what another individual knows from what they see. The test subjects were given the opportunity to steal a grape from a human competitor out of one of two containers. Both of the containers had twenty small jingle bells attached to the lid, but one of the containers had the ringers removed from the bells. The altered jingle bells produced no noise as opposed to the unaltered jingle bells that rang when the container was moved. The experimenters predicted that the subjects would reliably take the grape from the silent container, i.e.: the container with the altered jingle bells (Santos, et al. "Rhesus monkeys," 1177).

---

[25] Gaze-following occurs when an individual detects that another individual (either a conspecific, or a member of a different species) is not directed towards them and follows the line of other's sight to a point in space. It differs from joint-visual attention in that in the latter there is a focus of attention, i.e. the gaze is directed at a specific object.

[26] Laurie R. Santos, Aaron G. Nissen, and Jonathon A. Ferrugia, "Rhesus monkeys, *Macaca mulatto*, know what others can and cannot hear," *Animal Behavior*, vol. 71, 2006, pp. 1175-1181.

The experimenter began each trial by opening the lid of the container to his left, removing the grape inside, displaying it to the subject, and then returning it to its container, shaking the lid to display its auditory properties throughout. The experimenter then repeated this process with the container on his right. After the contents and auditory properties of the two containers had been displayed to the subject, the experimenter stood up and moved backwards away from the containers, squatted down, and placed his head between his knees. From this position the experimenter could not see the subject. The subjects were then allowed one minute to approach and visibly touch one of the two containers. The rhesus monkeys tested reliably chose the silent container over the noisy one. As Santos, Nissen, and Ferrugia note,

> Thus, subjects reliably picked the container that did not alert the experimenter to the fact that the grape was being removed. This result suggests that monkeys may take into account how auditory information can change what the experimenter knows. (Santos, et al., "Rhesus monkeys," 1178)

They acknowledge that an alternative explanation for their results may be that the subjects avoided the noisy container because they were afraid of it. In order to correct for this possibility they ran a second experiment in which the sound produced by the containers no longer influenced the experimenter's knowledge state. The design of this second experiment was identical to the first except that when the experimenter retreated from the containers he continued to look in the direction of the subject. In the second experiment only five of the sixteen subjects chose the silent container, providing evidence that the subjects did not intrinsically prefer the silent

container to the noisy one. Santos, Nissen, and Ferrugia concluded that the monkeys preferred the silent container only when the experimenter was not looking and lacked visual access to the monkey's approach.

The results from the studies conducted by Santos and colleagues reveal that monkeys "seem to reason about how seeing affects the relevance of hearing and not hearing" (1179). As stated by Santos, Nissen, and Ferrugia, "Primates' success in competitive tasks seem to come in contrast to poor performance on a number of mind-reading tasks that do not require competition" (1180). According to the researchers, the results from these studies indicate that a non-ape species is able to read mentalistic information from two perceptual modalities (vision and hearing) and to put this information together in order to determine how their behavior will affect what a competitor knows.

Advancements in experimental design have resulted in a dramatic change of position by many researchers. The later ToM studies involving competitive, rather than cooperative, tasks are assumed to more successfully replicate the environmental and social conditions in which such mindreading capacities likely evolved. The evidence provided by the more recent experiments involving competitive ToM tasks suggests that some non-human primates, both apes and monkeys, possess some of the capacities involved in mindreading. But sufficient evidence has not yet been provided for the possession of a full-blown, human-like theory of mind in any non-human primate species. The impasse reached between Povinelli et al. and Tomasello and Call remains. But Santos and colleagues work begins to expose the role of

presuppositions about the expression of ToM in humans (e.g. that it has some basis in empathy), which inform and constrain our inquiries into the presence or absence of that trait in other animals.

II. Parsimony (revisited)

Questions posed by Premack and Woodruff (1978) regarding non-human primates' possession of "mindreading" capacities remain controversial almost 40 years later. With respect to the capacity to engage in reasoning about the visual perspective and perceptual awareness of others, both those researchers who wish to attribute as well as those who wish to deny such capacities to nonhuman primates invoke 'simplicity' considerations in defending their claims. The interlocutors in the debate do not deny the validity of each other's data, but rather the explanation provided for those findings.

For example, many comparative psychologists investigating nonhuman primates mindreading capacities have focused on visual perspective taking; i.e. how seeing affects knowing and how knowing influences doing.[27] Because mental contents of a conspecific are not directly observable, researchers and theoreticians assessing another animal's capacity to infer a conspecific's mental states can only gain indirect access to their research subject's indirect access to another's mental state.

[27] Povinelli, for example, has focused his research on chimpanzees' ability (or lack thereof) to assume the visual perspective of a conspecific or human researcher.

Whereas mindreading involves higher-order cognitive processing (i.e. the possession of mental states about mental states), behavior-reading requires only first-order reasoning about non-mental phenomena. Povinelli and Eddy's (1996) finding that chimpanzees show no preference in begging for food from a human experimenter who can see them as opposed to one who cannot see them appears to provide evidence against mindreading in chimpanzees. But this conclusion has been challenged on methodological grounds. Because chimpanzee social life is dominated by competition over food, not communicative-cooperation, some have argued that the chimpanzee subjects' failure to perform well on such a task may be due to their not properly understanding the task itself, not the absence of a capacity to reason about others' perceptions. In an attempt to address this methodological problem, Hare, Call, and Tomasello developed a more "naturalistic" (i.e. ecologically valid) approach to the investigation of chimpanzees' understanding of visual perspective by utilizing a competitive rather than a cooperative task.

Hare et al. (2000) conducted a series of experiments in which they pitted a subordinate against a dominant chimpanzee in a food item selection task. Because dominant chimps tend to take all of the food available to them and punish subordinates who challenge them, the researchers predicted that if the subordinate is capable of reasoning about the dominant's visual perspective then they will prefer to take food that the dominant cannot see. Controls that rule-out various behavior-reading explanations included providing the subordinate with a head start (ruling out the hypothesis that the subordinate simply went for the food the dominant did not go

for) and replacing the opaque barrier occluding the dominant's vision with a transparent one (ruling out the hypothesis that the subordinate simply preferred food behind a barrier). Hare et al. take their findings to provide evidence that chimpanzees are sensitive to what others can and cannot see. Flombaum and Santos (2005) produced similar results with rhesus monkeys, suggesting that mindreading is not limited to apes.

Writing in 2008, it is clear that Povinelli and colleagues (2006, 2008) are and will remain unconvinced. They claim that all of these studies are *in principle* incapable of providing any evidence at all for mindreading and that the data presented can be sufficiently explained by attributing to the subjects a capacity for behavior-reading alone. According to Povinelli and Vonk, since mental states are not directly observable, subordinates can only infer particular mental states in dominants based on perceptually accessible features of the situation that they can observe. Povinelli and Vonk claim that purely first-order explanation can account for the positive results: subjects could reason according to a learnt behavioral rule which tells them that dominants are less likely to take food if an opaque barrier stands between their eyes and the food. Because Hare et al.'s experiments did not control for this possibility, Povinelli and Vonk propose that they are completely irrelevant to the question of whether nonhuman primates possess a theory of mind. Hare et al. admit that Povinelli and Vonk's first-order explanation is not ruled out by their control conditions, but they insist that such an account is extremely ad hoc. For each of the experiments performed a different behavior rule must be postulated.

Two different kinds of simplicity considerations are in play in the primate mindreading debate: simplicity as psychological unity and simplicity as parsimony of mental representation. According to Simon Fitzpatrick (2009),

> [Hare and colleagues] argument has been compared with a common simplicity argument against behaviorism: when patterns of behavior become very elaborate it is often simpler to ascribe sophisticated cognitive processes to organisms rather than having to postulate an enormous web of learnt associations between individual stimuli and responses. (270)

This anti-behaviorism argument is a form of the familiar 'poverty of the stimulus' argument provided for the Language of Thought Hypothesis according to which it is hard to account for how the subjects could have acquired the requisite associations through either their individual development, or the phylogenetic development of the species, to explain apparently novel behavior. According to Elliot Sober (2009),

> [A] model that postulates only lower-order intentionality (using *n* parameters to do so) is better than a model that postulates both lower-order intentionality (using *n* parameters) and higher-order intentionality (using *m* additional parameters) if the two models fit the data equally well. However, if introducing higher-level intentionality permits one to have fewer parameters overall while still fitting the data equally well, parsimony will speak in favor of introducing higher-level intentionality. (248)[28]

A unified model applies the same explanation to multiple data sets while a disunified model applies a different set of explanations to each data set. Because a unified model is more parsimonious than a disunified model, if they fit the data equally well, then the unified model can be expected to have greater predictive accuracy. A disunified

[28] Elliot Sober, "Parsimony and models of animal minds." *The Philosophy of Animal Minds*. Cambridge University Press, 2009, pp. 237-257.

model will have more adjustable parameters.

In the case of a series of experiments conducted by Hare et al. (2000, 2001) attempting to determine whether chimpanzees have the concept of "seeing," the data allows for a disunified theory that provides a different first-order explanation for each of the experiments' results, but a single first-order explanation that works for all the findings is difficult to formulate. A unified theory that attributes both first- and second-order intentionality can account for the entire body of data. Fitzpatrick has provided compelling reason to see the stagnation of the behavior-/mind-reading debate as arising not only from disparate conceptions of the principle of parsimony, but also from different presuppositions about the animal subject's individual learning history, or the evolutionary history of the species to which they belong.

VIII. Future directions

The push for ecological validity in attempts to experimentally demonstrate complex cognition in nonhuman animal subjects is admirable, and the positive findings provided by Hare and Santos are extremely informative about what monkeys know about what their conspecifics see. But the dissenters in the animal minds debate are not swayed. They have an endless supply of disunified associationist accounts of behavior that appear to others to be better explained by a unified cognitivist explanation.

In the primate mindreading debate, as it has played out, "positive findings" are in principle incapable of providing evidence because learning history could explain

results by virtue of the fact that the object of study, if it exists, is presumed to be a species-typical trait. But these researchers are not interested in ecologically invalid experimental paradigms because it says nothing of what a 'species' does. Because of this, we should explore the epistemic virtue that a non-naturalistic (i.e. ecologically invalid) experimental paradigm could afford in resolving the dispute over whether advanced forms of cognition are limited to the human animal. The move towards more naturalistic experimental paradigms has unfortunately led to a further stagnation in the association/cognition debate. Utilizing naturalistic experimental paradigms give the associationist an in to make the claim that the skill has a prior reinforcement history, whether onto- or phylogenetic.

We should overcome the restricted lens of inquiry that looks only to demonstrations of presumed species-typical traits of nonhuman animals that are proposed by the cognitivists to depend on some form of metacognition, and we should also move away from the primate literature. A research program initiated by Ronald Schusterman and continued today by Colleen Reichmuth at the University of California, Santa Cruz's Long Marine Laboratory does just that and will be provided as a case study in the following chapter.

Whereas all sides of primate mindreading debate assume a need for ecological validity, the research program on marine mammal cognition carried out at the Long Marine Lab over the last 30 years has instead focused on attempting to elicit non-species typical traits in their sea lion subject by teaching her to respond to certain relations among particular abstract symbols and then testing how she responds to

novel relations of those symbols. It is interesting to note that much of this work was carried out throughout the 1980s and that the findings reported in 1993 have never been picked up in the mainstream debate over animal cognition. This could be due in part to the primatocentric assumptions of many working in the field, both philosophers and researchers, but also due to the marine mammal researchers' explicit avoidance of naturalistic paradigms.

Ch. 3: Indeterminacy of Animal Minds: abstract reasoning in a non-human animal in an ecologically invalid setting

I. Introduction

Evolutionary psychologists investigating the cognitive mechanisms underlying human rationality have proposed that most, if not all, reasoning and decision making in their human subjects is carried out by domain-specific mental modules. They describe these modules as cognitive adaptations that have been refined by natural selection to assist our evolutionary forbearers in dealing with the types of adaptive problems they would have encountered in their physical and social environments. Richard Samuels and Stephen Stich have objected to this hypothesis of 'massive modularity' and have proposed instead that human reasoning and decision making is carried out in two different ways. Samuels and Stich present the claim that reasoning in humans is underwritten by two distinct neuronal systems, one of which is constituted by a number of domain-specific mental modules and another that is constituted by a set of domain-general problem-solving skills. In this chapter I will argue that Samuels and Stich's 'middle-way' is better able to account not only for the body of experimental data coming from studies of rationality in human subjects, but also for the results from studies of marine mammal cognition. Ronald Schusterman and colleagues' demonstration of the eventual acquisition of an equivalence concept by their sea lion subject, Rio, will be provided as a case-study for the attribution of a dual-processing cognitive system to a nonhuman animal.

II. Evolutionary psychology's biological approach to human rationality

Evolutionary psychology is an interdisciplinary field encompassing insights from both evolutionary theory and those of the social sciences with the aim of identifying the species-typical cognitive characteristics of humans. Evolutionary psychologists pursue this goal by applying the tools of the biological sciences to the study of the architecture of the human mind; they provide adaptationist accounts of our cognitive capacities.

Prior to the emergence of evolutionary psychology as a distinct discipline it was generally accepted that the mind contains physically locatable and functionally specialized circuits for the different modes of perception (sight, sound, smell, taste), but other cognitive functions like learning, reasoning, and decision-making were thought to be accomplished by very general-purpose brain circuitry (Fodor, 1983). Psychologists working in the heuristics and biases tradition, for example, assume that the mind consists predominately of a small number of general-purpose, content-independent mechanisms (Cosmides and Tooby, 3). These domain-general mechanisms are presumed to operate uniformly, regardless of context or content. In their discussion of the heuristics and biases tradition, evolutionary psychologists Leda Cosmides and John Tooby state, "The flexibility of human reasoning—that is, our ability to solve many different kinds of problems—was thought to be evidence for the generality of the circuits that generate it" (Cosmides and Tooby, 8).

Because the general-purpose problem-solving strategies posited by the heuristics and biases researchers contain no domain-specific knowledge, these

reasoning procedures should not be restricted to operating on input of a particular class. That is, they should be able to support inferences regardless of the specific content. But evolutionary psychologists have proposed that different adaptive problems require different problem-solving strategies, so problem-solving skills will be, in many cases, domain-specific. That is, evolutionary psychologists claim that the special-purpose, content-dependent problem solving skills they have purportedly identified through their experimental manipulations are specialized solutions developed by natural selection for assisting us to cope with particular adaptive problems that were faced by our evolutionary forbearers.[29]

Evolutionary psychologists view the human mind as a set of information-processing mechanisms designed by natural selection to solve adaptive problems faced by ancestral populations. They question the search for a limited number of *domain-general* 'judgment under uncertainty'[30] problem-solving skills and propose instead that we should investigate *domain-specific* cognitive mechanisms that would have assisted our evolutionary ancestors in solving the types of adaptive problems they would have faced. So, for example, instead of evaluating their subjects' ability

---

[29] Leda Cosmides and John Tooby, "Evolutionary Psychology: A primer," Center for Evolutionary Psychology, University of California Santa Barbara, 1997. (Available online at http://www.psych.ucsb.edu/research/cep/primer.html).
For a contrary position, see Jerry Fodor, *The Mind Doesn't Work That Way: The scope and limits of computational psychology*, MIT Press, 2001.

[30] See Amos Tversky and Daniel Kahneman, "Judgment under uncertainty: Heuristics and biases," *Science*, vol. 185, 1974, pp. 1124-1131. In this paper Tversky and Kahneman propose three domain-general heuristics that humans employ when making a choice under conditions of uncertainty: representativeness (employed when judging the probability that a given event or object belongs to a particular class or process), availability of instances or scenarios (employed when assessing the frequency of a class or plausibility of a certain development), and adjustment from an anchor (employed when making a numerical prediction in the absence of knowledge of a relevant value). They propose that although these heuristics are highly economical and, in general, effective, they do lead to systematic and predictable errors.

to solve content-neutral reasoning problems, they instead provide their subjects with highly contextualized tasks that mimic problems that would have been faced in the real lives of our ancestral populations, like those involved in social exchange situations.

III. Modularity of mind

Richard Samuels (2000) has distinguished two notions of modularity appearing in the literature: Chomskian modules, which are understood as domain-specific bodies of mentally represented information that are inaccessible by other domains of reasoning, and computational modules, which are taken to be domain-specific computational mechanisms that only operate on input from a particular domain (Samuels, 2000, 14). Samuels identifies the modules invoked in the scientific literature as the latter type. That is, when cognitive scientists use the term 'module' they are almost always referring to mental structures that are invoked to explain cognitive capacities and it is usually assumed that modules are functionally specific in the sense of being dedicated to solving a restricted class of problems in a restricted domain (Samuels, 2000, 16).[31] For example, Cosmides and Tooby refer to the modules they postulate as "functionally dedicated computers," i.e. domain-specific computational mechanisms (Cosmides and Tooby, 1995, xiii; quoted in Samuels,

[31] Max Coltheart (1999) has also identified this distinction in the literature, referring to the notion of modularity invoked by evolutionary psychologists as involving "processing modules" and that discussed by linguists (among others) as "knowledge modules" (118). Like Samuels and Stich, Coltheart's concern is with the former type, but unlike Samuels and Stich, Coltheart acknowledges that "processing modules will, in general, incorporate knowledge modules—[for example,] the syntactic processor will have, as part of its internal structure, a body of knowledge about syntax" (118).

2000, 20). According to Cosmides and Tooby, "We have all these specialized neural circuits because the same mechanism is rarely capable of solving different adaptive problems" (Cosmides and Tooby, 8).[32]

Although Jerry Fodor (1983) introduced the concept of the modularity of mind, he did not propose to have offered either a precise definition of 'modularity' nor to have provided any necessary conditions that a cognitive system must obtain for the term 'module' to accurately be applied. Rather, Fodor offered five features of a cognitive system that will typically be present, to some degree, in those systems that can be identified as modular in structure. The five characteristic features of modularity put forth by Fodor are: domain-specificity, innateness, neural specification (i.e. they are 'hardwired' in the brain), autonomy, and not assembled[33]

---

[32] On a strictly individual-level model of selection, evolutionary biologist Robert Trivers (1971) has demonstrated that cooperation could not stabilize in a population if the members of that population did not possess a cognitive mechanism that allows them to identify freeriders in the group (in addition to an ability to remember who those freeriders are). In response, evolutionary psychologist Leda Cosmides (1989) has postulated the existence of a Cheater Detection Module (CDM), an individual's possession of which would solve the problem of freeriders in a population whose members exhibit selfless behaviors by providing a mechanism for the identification of those freereiders (i.e. conspecifics that commit violations of the social contract). The CDM is proposed to take social exchange contracts in the form of prescriptive conditional statements (e.g. "*If someone is eating from the kill, then they must have participated in the hunt*") as its input and to not respond to other kinds of conditionals, even if they share the same logical form (e.g. "*If it rains, then the streets will be wet*"). To know if someone has violated the social contract, one needs to know if they took the benefit and whether or not they have met the requirements entitling them to that benefit. So, to identify a violation of the prescriptive conditional rule, "*If someone is eating from the kill, then they must have participated in the hunt*", one must look to the individuals eating from the kill and confirm that they participated in the hunt as well as to those who did not participate in the hunt and confirm that they are not eating from the kill.

[33] Although Fodor (1983) included 'not assembled' in his preliminary discussion of the characteristic features of modular systems, he provides no reason for this proposal nor does he include it in his detailed discussion of the features of modular systems that comes later in the text. Fodor does posit that in many cases there will be interlevels of representation within a module and that each of these interlevels will have a further domain to which it is specific. For example the language module will have within it a phonetic interlevel as well as a lexical interlevel which are each specific to their respective domains (Fodor, 1983, 137). In other words, even on Fodor's own account a modular system can be decomposed into smaller, domain-specific subsystems. Max Coltheart contends that

(Fodor, 1983, 37). But Max Coltheart (1999) has contended that on a further elaboration of Fodor's arguments, the following definition of 'modularity' emerges: a cognitive system is modular when and only when its application is domain-specific (Coltheart, 118). Coltheart defines 'domain-specificity' as "not responding to inputs except those of a particular class" and defends the claim that domain-specificity is the feature at the heart of the concept of modularity (Coltheart, 119).

Coltheart provides an example of the empirical identification of a domain-specific module for face-recognition, i.e. "a cognitive system that responds when its input is a face, but does not respond when its input is, say, a written word, or a object, or someone's voice" (118). If we begin with the presumption that there is a single, domain-general cognitive system responsible for visual recognition, then we should expect it to accept input from a variety of different types of sources (e.g. faces, objects, and written words). But neurophysiological studies have demonstrated that there are patients with impaired visual word recognition who retain their facial recognition skills (Miozzo and Caramazza, 1998) as well as patients with impaired visual word recognition who retain their visual object recognition skills (De Renzi and Di Pellegrino, 1998). Additional neurophysiological studies have identified patients with impaired visual object recognition who retain their visual word recognition (Rumiati and Humphreys (1997) as well as patients with impaired face recognition who retain their visual object recognition (Buxbaum et al. (1996).

'assembled' in regards to modular systems means simply that "a module itself can have an internal modular structure" (118).Ned Block ("The Mind as the Software of the Brain." *Thinking: An invitation to cognitive science*, " edited by Smith and Osherson , 1995) has proposed that if we accept that modular systems can be decomposed into a set of smaller modules, then we have no reason to reject assembled as a feature of some modular systems.

According to Coltheart, "This collection of results refutes our original idea that there is a cognitive module whose domain is the recognition of all forms of visual stimuli. Instead, it suggests that there are three separate modules: a face-recognition module, a visual-object-recognition module, and a visual-word-recognition module" (119), with the respective domain-specific input classes of faces, visual objects, and printed words. According to Coltheart, if a cognitive capacity is elicited in a variety of domains (i.e. response is not restricted to stimuli of a particular class) it cannot be a 'module' on the Fodorian account of modularity.

Fodor (1983) reminds us that cognitive capacities are functionally, rather than physiologically, identified (Fodor, 1983, 98). Nevertheless, each of the sensory input systems (sight, sound, smell, taste), as well as language, do appear to have a characteristic neural architecture associated with them. But, as Fodor points out, "There is, to put it crudely, no known brain center for *modus ponens*" (Fodor, 1983, 98). So, in terms of the modularity of the peripheral systems, there is evidence of physical correlates in the central nervous system, but not for those modules postulated by the evolutionary psychologists. It is important to bear in mind that the failure to demonstrate the neuroanatomical modularity of a cognitive system would not, by itself, provide evidence against the cognitive modularity of that system (Coltheart, 119).[34] It follows that evolutionary psychologists postulating the modularity (i.e. domain-specificity) of a cognitive system like learning, reasoning, or decision-

[34] A cognitive system could depend on the activation of a specific constellation of multiple brain regions and still not be accessible by other cognitive systems that activate elements of those same regions, or even by all of the neural circuitry within those regions that are activated.

making need not be dissuaded by neurophysiologists' failure to identify the neuroanatomical modularity of such systems.

Whereas some other approaches to psychology have focused on how we solve those problems we are bad at (namely, the heuristics and biases tradition), like learning math or playing chess, evolutionary psychologists have focused on the ones that our evolutionary history has made us good at, like those involved in social interactions. Evolutionary psychologists begin by identifying specific adaptive problems that would have been faced by our evolutionary ancestors and then design and conduct experiments aimed at determining whether or not we do in fact posses those specialized reasoning skills, expecting to find disparate results from subjects' performance on the formally identical but context-independent and content-neutral tasks examined by the heuristics and biases researchers.

IV. Assessing human rationality

In order to assess a human subject's ability to solve reasoning problems, psychologists from both the heuristics and biases tradition as well as those working from an evolutionary approach have designed and conducted a number of experiments known collectively as 'selection tasks.' For example, in an experimental paradigm known as the Wason 4-card selection task, the subject is asked to identify violations of a conditional rule, i.e. a rule of the form *If p, then q* (Wason, 1966; Wason and Johnson-Laird, 1972). When heuristics and biases researchers performed this test, the problem was presented as follows:

> Here are four cards. Each of them has a letter on one side and a number on the other side. Indicate which of these cards you have to flip over in order to determine whether the following claim is true: 'If a card has a vowel on one side, then it has an odd number on the other side.'
>
> **E C 5 4**
>
> (Samuels and Stich, 281)

Evolutionary psychologists retained the logical structure of the task, but adapted the content so as to provide the subjects with a more realistic problem (i.e. one that they presume our evolutionary history would have equipped us with a method for dealing with):

> You are a bouncer in a Boston bar, and you will lose your job unless you enforce the following law: 'If a person is drinking beer, then he must be over 20 years old.'
> One side of a card tells you what a person is drinking and the other side tells that person's age. Indicate only those card(s) you definitely need to turn over to see if any of these people are breaking the law.
> **Beer Coke 25 16**
>
> (Samuels and Stich, 291)

In order to determine if the conditional rule (*If p, then q*, or, $(x)\ (Px \supset Qx)$) has been violated, one should turn over the cards that represent the values *p* and $\sim q$ (i.e. not *q*) because it is only in the case that *p* is true and *q* is false that the conditional is violated. In the first formulation the conditional statement *'If a card has a vowel on one side, then it has an odd number on the other side'* indicates that we must look at the card marked with a vowel (i.e. '*E*'), to determine whether it has an even number on the other side, as well as the card representing the negation of *'odd number*' (i.e. '*4*'), to see if it has a vowel on the other side, either of which would be violations of

the conditional statement. In the second formulation, the conditional statement is *'If a person is drinking beer, then he must be over 20 years old.'* In this case one would need to look at the card marked *'Beer'* to see if the other side is marked '*16*' and the card representing the negation of *'over 20 years old'* (i.e '*16*') to see if it is marked '*Beer*.' Despite the fact that the same logical rule is involved in both formulations of the selection task, whereas subjects perform poorly on the former formulation, averaging only about 25% correct responses, about 75% of subjects get the right answer with the latter formulation.

Prior to the evolutionary psychologists' reformulation of the Wason 4-card selection task, the findings presented by the heuristics and biases researchers appeared to reveal the fundamental and systematic irrationality of human decision making processes. Cosmides and Tooby acknowledge that the findings reported in the heuristics and biases literature indicate that people are often not very good at detecting violations of *if, then* rules, but have proposed that this shows that the Wason selection task provides researchers with "an ideal tool for testing hypotheses about reasoning specializations designed to operate on social conditionals, such as social exchanges, threats, permissions, obligations, and so on" … "By seeing what content-manipulations switch on or off high performance, the boundaries of the domains within which reasoning specializations successfully operate can be mapped" (Cosmides and Tooby, 20). That is, according to Cosmides and Tooby, by manipulating the content of the reasoning task (in this case, the content of the conditional rule *if p, then q*), researchers can experimentally evaluate performance in

different contexts and can thereby identify the boundaries of these proposed domain-specific reasoning skills.

The evolutionary psychologists' claim that reasoning skills are, in many if not all cases, domain-specific thereby gains support not only from their own findings but also from the heuristics and biases researchers findings of apparent systematic irrationality. In providing their subjects with non-contextualized reasoning problems the heuristics and biases researchers failed to elicit their subjects' content-dependent problem-solving skills. Evolutionary psychologists take these apparent failures of human reasoning as evidence that many if not all reasoning skills are specialized solutions to specific problems faced by our evolutionary ancestors that are only "switched on" in the particular contexts that they provided adaptive value to our evolutionary forbearers.

The Wason selection task is just one of many experiments evolutionary psychologists have designed and conducted in their attempt to demonstrate that most, if not all, reasoning and decision making in humans is subserved by normatively unproblematic 'elegant machines' that operate both efficiently and effectively, albeit only in those types of social situations in which they were evolved, and that worries of systematic irrationality presented in the heuristics and biases literature are thereby unfounded. A clear conflict between the alternate pictures of rationality provided by the two traditions is apparent: "[M]any observers have concluded that the view of the mind and of human rationality proposed by evolutionary psychologists is

fundamentally at odds with the view offered by proponents of the heuristics and biases program"(Samuels and Stich, 296).

V. Samuels and Stich's "middle-way"

Samuels and Stich provide what they take to be the appropriate conclusion to draw about ordinary human rationality when the entire body of experimental data is taken together. They propose that we should be neither as pessimistic as the heuristics and biases account indicates nor as optimistic as the evolutionary psychologists. Rather, they conclude, "People do make serious and systematic errors on many reasoning tasks, but they also perform quite well on many others. The heuristics and biases tradition has focused on the former cases, while evolutionary psychologists have focused on the latter" (Samuels and Stich, 296).

The 'middle-way' that Samuels and Stich offer displays the compatibility of the two approaches. It gains support from a family of 'dual-processing' theories, according to which, "reasoning and decision making are subserved by two quite different sorts of systems. One system is fast, holistic, automatic, largely unconscious, and *requires relatively little cognitive capacity*. The other is relatively slow, rule based, more readily controlled, and *requires significantly more cognitive capacity*" (Samuels and Stich, 297, emphasis added). Whereas many of the tasks studied by evolutionary psychologists fall into the category of the former, most of those tasks examined in the heuristics and biases tradition fall into the latter category. So, dual-processing theorists like Samuels and Stich propose that a specialized, domain-

specific reasoning system will likely be at work in the types of highly contextualized reasoning tasks examined by evolutionary psychologists (e.g. social exchange situations) and that a non-specialized, domain-general reasoning system will be at work in the less contextualized reasoning tasks examined by heuristics and biases researchers.

Samuels and Stich have revealed the narrow-mindedness of theories of cognition that attempt to account for all of human reasoning as strictly domain-general or as solely domain-specific skills.[35] The dual-processing explanation of human reasoning attempts to correct for these inadequacies in accounting for the disparate experimental results by providing an account that accepts that reasoning in humans is, in some cases, carried out by the general purpose skills that are utilized in diverse contexts and across varieties of content and, in other cases, underwritten by

[35] Although the proponents of evolutionary psychology's tenet of massive modularity have for the most part stuck to their guns rather than accepting Samuels and Stich's proposal of a dual-processing cognitive system underlying human reasoning, others who are less invested in the outcome of the dispute have provided neuroanatomical accounts that support the dual-processing theory over either of its primary alternatives. For example, see Michael Anderson's "Neural reuse: A fundamental organizational principle of the brain," *Behavioral and Brain Sciences*, vol. 33, 2010, pp. 245-313. And for how dual-processing accounts apply to issues of continuity between the cognition of humans and other animals, see Linda Hermer and Elizabeth Spelke, "Modularity and development: the case of spatial reorientation," *Cognition*, vol. 61, 1996, pp. 195-232.
Daniel Kahneman, who together with Amos Tversky (1937-1996) founded the heuristics and biases approach to understanding human decision making under conditions of uncertainty, has been persuaded by the dual-processing theorists. Kahneman was awarded the Nobel Prize in Economics in 2002 for establishing a cognitive basis for human errors that arise as a result of our reliance on heuristics and biases when making economic decisions. In 2011 Kahneman published *Thinking, Fast and Slow,* which summarized his and Tversky's research from the last 3 decades of the 20th century, but presented it alongside a dual-processing approach to understanding human rationality. The book examines the dichotomy between what he refers to as System 1, which is fast, instinctive, and emotional, and System 2, which is slower, more deliberate, and more logical. Kahneman's focus in this text remains on the cognitive biases systematically associated with each type of thinking, but his acceptance of a dual-processing approach is central to his thesis that humans are far less rational than we are accustomed to believe. This is because System 2 is highly distractible and difficult to engage, so as a result we often fall back on the emotion-based, unconscious, snap judgments of System 1 despite believing that we have made a conscious, deliberate choice.

specialized reasoning skills that are activated only in particular contexts involving problems with specific content. I propose that just as the single-minded approach that seeks to explain all of human reasoning either by the heuristics and biases' domain-general account or by the evolutionary psychologists' domain-specific account is incapable of explaining the diversity of human reasoning, so too is an account that seeks to explain all of nonhuman animal problem solving by reference solely to context-independent and content-neutral cognitive capacities.

VI. Reasoning in non-human animals

In their introduction to a special issue of *Animal Cognition* dedicated to addressing the question of "Animal Logics," Watanabe and Huber describe a variety of experimental evidence coming from studies of animal cognition that indicate that at least some aspects of animal cognition are logical. They state, "On the basis of Darwinian evolution, it seems reasonable to assume that specifically human abilities are not entirely unrelated to those of our animal relatives, and of course there are many aspects of human behavior that are by no means specifically human" (Watanabe and Huber, 236). Watanabe and Huber accept the evolutionary psychologists' claim that "logical thinking is a result of natural selection or phylogenetic contingency[36]," as well as that, "the logical reasoning of humans may be

[36] 'Phylogeny' refers to the developmental evolutionary history of a species, i.e. the species history of an organism. Phylogenetic contingencies are a species' history of selection pressures in the course of evolution. In seeking to determine the source of an organism's behaviors, it is useful to distinguish between these phylogenetic contingencies of a species and the ontogenetic contingencies of a particular organism (i.e. that organism's reinforcement history, or, anything that has increased the likelihood that a particular response will occur under certain conditions). See Jerry A. Hogan, "The

based on some unique cognitive modules and distinct educational environments" (Watanabe and Huber, 238), but they also acknowledge that, "Many animals enjoy adaptation without advanced forms of cognition, by simply using 'rules of thumb' to cope efficiently with problems they face in real life" (Watanabe and Huber, 239).[37]

Unlike genetically preprogrammed behavior, cognitive behavior allows for flexibility of response. The ability to learn by trial-and-error provides a familiar example of a rule of thumb, or rough heuristic, which does require some level of cognitive processing. It results in a change in behavioral response to a certain stimulus in light of prior negative and/or positive reinforcement incurred from previous responses to that stimulus. Negative reinforcement (i.e. failure) conditions the organism to refrain from such behavior in the future and positive reinforcement

---

structure versus the provenance of behavior," in *The Selection of Behavior: The Operant Behaviorism of B.F. Skinner*, edited by A. Charles Catania, Cambridge University Press, 1988, pp. 433-436.

[37] It is essential to note that the lower-level processing that gives rise to the use of rules of thumb is not non-cognitive, i.e. instinctual, but it is not a case of advanced cognition. In regard to non-cognitive behavioral mechanisms, E.O. Wilson (1975) provides the following intuitive definition: "An instinct, or innate behavior pattern, is a behavior pattern [i.e. fixed-action-pattern], that either is subject to relatively little modification in the lifetime of the organism, or varies very little throughout the population, or (preferably) both" (Wilson, 26). The behavior of the female *Sphex* wasp provides an informative example of an organism engaged in such a non-cognitive, stimulus-bound, instinctual response. When the time comes to lay her eggs, the *Sphex* wasp displays behavior that may appear rational to the naïve onlooker, but lacks the characteristic feature of cognition, i.e. flexibility of response. According to Wooldridge (1968), the *Sphex* wasp engages in a highly ritualized species-typical pattern of behavior. This routine consists of digging out a burrow, locating and stinging a cricket, bringing the paralyzed cricket to the threshold of the burrow and then going in to confirm that all is well inside. The wasp then emerges from the burrow and drags the cricket inside before laying her eggs. But, as Wooldridge demonstrated, if the cricket is moved a few inches away from the threshold while the wasp is inside, the wasp will bring the cricket back to the threshold and will repeat the preparatory process of checking that all is aright inside before bringing the cricket in. When Wooldridge continued to move the cricket away from the threshold every time the wasp entered the burrow, she was caught in an endless cycle of repetitious behavior (D. Wooldridge, *Mechanical Man: The Physical Basis of Intelligent Life,* 1968, quoted by Dennett, 1984, p. 11). Also regarding the nesting behavior of the *Sphex* wasp, E. O. Wilson wrote, "In not the remotest sense is learning involved in such a machinelike response" (Wilson, 27). This utter lack of flexibility in behavioral response indicates that cognitive processes are not at work in guiding the wasp's behavior and that, rather, the wasp is engaged in a stimulus-bound fixed-action-pattern.

(i.e. success) conditions the organism to repeat the behavior in similar circumstances in the future.

An organism's ability to learn by trial-and-error indicates that more is at play than inflexible stimulus-bound responses (i.e. fixed-action-patterns) because it requires the organism to alter their behavior in light of information acquired through experience. In other words, it involves flexibility of response to a given stimuli. Yet it is still a very general problem solving strategy which roughly consists of the following rule: *if some method of obtaining your goal does not work, try something else and continue to do so until success is met. When a solution is found, use it the next time the problem is encountered*. Learning by trial-and-error is a rule of thumb because it is a problem solving strategy that applies in all possible contexts; the cognitive mechanisms underlying it are not restricted to input of a particular class. Association is another, even more basic, rule of thumb that is also not unique to humans and that plays a role in learning by trial-and-error. Associative learning is based on the following principle: *If A has been followed by B every time it has been encountered it in the past, then expect B when you encounter A in the future*. As with the cognitive mechanisms underlying the ability to learn by trial-and-error, the input for such a mechanism is not restricted to that of a particular class.[38] Rules of thumb (like trial and error and association) are cognitive but not abstract, i.e. they are not

[38] Both learning by trial-and-error as well as associative learning can be driven by operant conditioning, which involves voluntary behaviors and is maintained over time by the consequences that follow those behaviors. An attempt that results in success (a positive reinforcer) may be repeated in the future as a result of the reinforcement rather than due to insight learning as to why the attempt was successful. A failed attempt will be refrained from in the future because of the negative stimulus that resulted from the given behavior, rather than because of an understanding of why or how the attempt failed.

advanced forms of cognition because they cannot be deployed in a context that transcends the perceptual contingencies of the situation in which they were acquired. In contrast, fixed-action-patterns are non-cognitive because they cannot be altered in light of experience.

In order to determine if an organism is cognitive (at any level), what is called for is a testing paradigm in which the stimulus itself could not result in a fixed-action-pattern response, i.e. a genetically preprogrammed behavior that is completely inflexible and unresponsive to the presence or absence of reinforcement from the organism's environment.[39] With an ecologically valid (i.e. naturalistic) experimental set-up[40] it is much more difficult to rule out such a possibility because by mimicking the organism's natural habitat the response being evoked could be a reflexive behavior that conferred an adaptive advantage on the organism's evolutionary ancestors.[41] Additionally, in the case of demonstrations of advanced cognition, ecologically valid experimental designs cannot rule out the possibility that the organism has a prior reinforcement history to that stimulus which gave rise to a

[39] Reinforcement is defined by the effect that it has on behavior. Pavlov (1903) introduced the term to describe the strengthening of the association between an unconditioned stimulus and conditioned stimulus that results when the two are paired together. For Pavlov, the term denotes both the establishment as well as the strengthening of an association between a conditioned stimulus (ex. ringing bell) and its unconditioned parent stimulus (ex. salivating dog) (Pavlov, 1928). Currently, the term is used more frequently in relation to response learning rather than the stimulus learning that Pavlov was concerned with. This change in meaning of the term 'reinforcement' was inaugurated by Thorndike's *Law and Effect* (1911) and continued by Skinner's (1933) adoption of Pavlov's terminology to denote the strengthening of a stimulus-response association. For Pavlov, what was reinforced was the association between two distinct stimuli (S-S learning), but for those that came later the term primarily is used to refer to the strengthening of association between a stimulus and a behavioral response (S-R learning).

[40] As explained in the previous chapter, ecological validity in experimental design refers to the degree to which the experimental setting and the nature of the tasks put to the subject mimic the organism's ecology (i.e. natural habitat).

[41] See footnote 36 on the behavior of the *Sphex wasp*.

trained association between that stimulus and the response being sought. For example, a subordinate chimpanzee may approach food a dominant has not seen and avoid food a dominant has seen not because they recognize the relationship between seeing and knowing, but rather because doing so has been advantageous to them in the past. It is a main contention in this paper that ecologically invalid experimental set-ups provide more robust evidence that some behavior is the result of cognitive processes rather than a species-typical reflexive response to a given stimulus (i.e. a fixed-action-pattern) because they provide a novel context in which to examine the cognitive mechanisms employed by the subject.

Behavioral flexibility and the resultant ability to learn by trial-and-error provide evidence of cognition in a broad sense in that the subject is not restricted to fixed-action-pattern responses to the stimuli it is presented with. But demonstrations of learning abilities which depend only on lower-level cognitive processes do not provide evidence of the capacity for logical reasoning and the possession of abstract thought on which that reasoning depends, which involve the operation of more advanced cognitive mechanisms.

Piaget (1954) developed a nonlinguistic paradigm for assessing whether the capacity for logical reasoning has emerged in a human child. If the subject is able to derive a relationship between items that have never been presented together before, then logical reasoning has emerged. According to Watanabe and Huber, "reasoning involves computations over logical forms" (237), and, "the basic requirement for logical reasoning is a process called 'abstraction,' which is the identification of

regularities in the environment with the formation of inner models or representations" (Watanabe and Huber, 241).[42] What these definitions of abstraction share is the idea that the ability to abstract allows an organism to recognize the underlying form of concrete relations between particular stimuli that have been explicitly learned through interactions with one's environment and to transform those relations into formal rules of inference that can be applied to novel stimuli and the relations amongst them.

Abstraction enables an organism to avoid the slow process of learning via trial and error every new relation between objects and events in their environment and instead to be able to generalize from past experience to current particular instances of the same form. Watanabe and Huber state, "[A]cquiring the capacities to reason or think 'logically' in order to solve problems in their physical and social world more efficiently does not mean that subjects would lose their capacity to use associative processes or even rules of thumb. But the same is true for humans" (Watanabe and Huber, 243). In regard to the question of whether animals (or humans) are ever logical, it is important to keep in mind that rationality is not an all-or-nothing phenomenon. That is, the capacity for rationality does not require that one always be rational in every domain. If it was necessary for an organism to never exhibit irrationality in order to be considered a rational being, no human would qualify.

[42] Examples of logical forms include $p \supset q$, p, q, $\sim p$, and $\sim q$. An example of a computation over such logical forms is the rule of inference known as *modus tollens*, represented as $p \supset q$; $\sim q$; $\therefore \sim p$. The capacity for abstract reasoning enables an individual to recognize general rules encountered in their experiences with relations between stimuli and to then apply those rules to novel instances of the same form.

VII. Investigating the mechanisms of non-human animal minds

Ronald Schusterman and his colleagues at the Long Marine Laboratory at the University of California, Santa Cruz, have performed experiments designed to determine whether California sea lions possess the capacity to form abstract concepts and have published their findings in a number of psychological journals in addition to journals that specialize in marine mammal research, but their findings have been largely ignored in the philosophical debates over the status of advanced cognition in nonhuman animals. According to Kastak and Schusterman, abstract conceptual thought involves "the generalization of a particular problem-solving strategy on the basis of experience with a few examples of a particular problem" (Kastak and Schusterman, 1992, 414).

The ability to organize stimuli into classes allows an animal to generalize skills that they have learned in relation to a particular problem and to apply those skills more broadly in a variety of contexts. For example, in the case of social hierarchy, if an animal (human or otherwise) learns that a particular conspecific is dominant to them (bDa) and that another conspecific is dominant to that first conspecific (cDb), the ability to infer that the second conspecific is also dominant to them (cDa) would provide useful knowledge without requiring the expenditure of energy involved in engaging in a threatening social interaction with that conspecific in order to gain that knowledge directly. The transitive nature of such a relational property as dominance, the recognition of which may be a product of the operation of a cognitive mechanism that accepts input from the domain of social interaction, may

then be abstracted to a non-social context. Kastak and Schusterman state, "Specifically, abstract concepts such as *same/different*, relational concepts such as *larger than/smaller than*, and perceptual concepts such as *animal/non-animal* or *fish/non-fish* should allow an animal to increase its own fitness, by adapting rapidly to changing environmental conditions" (Kastak and Schusterman, 1992, 414).[43]

The match-to-sample (MTS) experimental paradigm is the most common procedure utilized in the lab to test a subject's ability to acquire concepts. In a simple visual MTS experiment, a subject is presented with a sample stimulus and is then given a choice between two comparison stimuli, only one of which matches the sample stimulus. The subject's choice of the correct (i.e. matching) comparison stimulus is reinforced. Many psychologists implicitly assume that success at a simple MTS test (selecting the comparison stimulus that 'matches' the sample stimulus) indicates that the subject is applying an abstract concept of 'sameness' or equivalence between the sample and comparison in choosing the correct comparison stimuli, but such an assumption is often under-determined by the experimental data. The subject's ability to form conditional relations between the sample stimulus and the matching comparison stimulus (demonstrated experimentally by correct responses) does not necessarily indicate that the subject takes the two stimuli to be equivalent to

[43] We must keep in mind that an ability to recognize food, for example, does not indicate that an organism possesses the concept 'food.' The question is really whether the animal is solely a biologically determined reflex machine—and this is extremely difficult to determine in ecologically-valid testing environments. If an animal only eats a restricted variety of food items we cannot determine whether they have a concept of 'food' because each of those food items could be an innate-releasing stimuli for the fixed-action-pattern response of consuming. The ability to recognize that something other than those typical items may be worth consuming and then learning to recognize and consume that food again when it is encountered in the future is a sign of flexibility of behavior and thereby of the operation of cognitive processes.

one another because the testing paradigm is not able to eliminate the possibility that the subjects may be relying on a general-purpose rule of thumb in making their choice. That is, the subject's choice could be the product of a cognitively more simple stimulus-bound response and not the result of a cognitively advanced abstract reasoning procedure. Success at a standard match-to-sample task could be explained by purely pictorial matching, not by reference to an understanding of the relations between symbolic representations such as are required when forming associations between, for example, a spoken word (or written word) and an object (or image of that object) on the first presentation. The space between stimulus-bound rules of thumb and abstract conceptual thought is not an vacant gap, but is populated by intermediary cognitive processes, including the capacity to reason by exclusion.

VIII. Reasoning by exclusion

The notion that some nonhuman animals possess the capacity for logical inference is not new to the literature, but has only recently seen a resurgence in interest. In the second century A.D. Pyrrhonian philosopher Sextus Empircus recounted a story told by Stoic philosopher Chrysippus 500 years earlier. According to Chrysippus, a dog who had chased an animal to a place where three streets intersect and having lost sight of the animal, smells the first and then the second road and then rushes off down the third road without checking it for the animal's scent. According to Chrysippus, as told by Sextus, the dog implicitly reasons as follows: "'the animal went either by this road, or by that, or by the other: but it did not go by

this or by that, therefore he went the other way" (Sextus Empiricus, 69-70). That is, the dog made a valid inference by the rule of *modus tollendo ponens* (also known as inference by exclusion). This rule can be represented as follows: *(A v B) v C; ~ (A v B); ∴ C* and is also referred to as *disjunctive syllogism*.[44]

The exclusion effect may occur in testing situations in which a novel (nonfamiliar) comparison stimulus is presented alongside a familiar comparison stimulus as the choices to be matched to a novel sample stimulus. The subject's reaction to the nonfamiliar sample is to select the nonfamiliar comparison. Choice of the correct comparison stimulus may lead to the appearance that an equivalence

[44] Recently, dogs' ability to reason by exclusion has garnered significant attention from the comparative cognition community. According to primatologist Josep Call, inference by exclusion "consists of selecting the correct alternative by logically excluding the other potential alternatives" (Call, 393). In 2007 Rico, a male border collie in Germany, gained worldwide attention after appearing on the cover of *Time* magazine. Researchers from the Max Plank Institute in Leibniz, Germany had published findings a 2004 issue of *Science* that dogs possess the ability to learn by exclusion. In humans, the ability to match a new word to an object that does not yet have a name is present at a very early age and is believed to play an essential role in vocabulary development. Juliane Kaminski and her colleagues at Max Plank demonstrated that Rico was able to acquire more than 200 words (specifically, object names) not only through trial and error, but also by making use of inference by exclusion. In the last couple of months Rico's success has been eclipsed by the findings that have made headlines in both the public literature as well as the scientific literature. When John Pilley, a retired psychology professor from Wofford College in South Carolina, heard of Kaminski's work with Rico he wondered how far such a training program could go. He acquired a female border collie from a local breeder, named her Chaser, and began a very intensive research program with the aim of demonstrating the vast extent to which a dog can acquire new associations via inference by exclusion. Over a period of 3 years Chaser learned the names of 1,022 objects. At the beginning Chaser relied on trial and error to learn object names, but once she had acquired a small repertoire of them to allow it, subsequent training was carried out in contexts that required her to make use of exclusion. According to Pilley and his co-researcher Ried, in the case of learning by exclusion, "This choice response cannot be based on associative learning mechanisms because the name and object referent are not presented together in a single trial" (9). Unlike the standard match-to-sample paradigm, both Kaminski and Piley utilized a fetching paradigm with Rico and Chaser in which the dog was asked to fetch a novel object from a selection of familiar ones and then later to fetch that object from a selection of both familiar and unfamiliar items to confirm that the object had been associated with the novel word. In order to experimentally demonstrate Chaser's ability to acquire object names via exclusion, the researchers placed a single novel object among seven objects that were familiar to Chaser in a room adjacent to the room the researchers would direct her from. Chaser was first asked to retrieve two familiar items by name and was then asked to retrieve the novel item by its novel name. Upon successful retrieval of the novel item she was rewarded by being told "Good dog" (Piley and Reid, 10).

concept has emerged, but the subject may be relying on exclusion to inform them that the familiar comparison stimulus is not the right choice rather than that the novel comparison stimulus is equivalent to the novel sample stimulus. Under these conditions the animal can respond correctly by excluding choice stimuli that have previously been associated with other samples. In such cases they are not matching, that is, they are not performing a conditional discrimination. But it is important to keep in mind that they are doing something interesting and that it does provide evidence of cognitive processing. For example, if a subject has learned to match ♣→♣ and ♦→♦ and is then introduced to ♥, in the context of a choice between ♥ and ♣ (or ♦), the correct choice of ♥ could be due to their response that the sample stimulus, ♥, is not the familiar ♣ (nor ♦), rather than due to recognition of a 'sameness' identity relation between sample ♥ and comparison ♥. The ability to reason by exclusion as demonstrated by success in a MTS task does provide evidence of cognition in the sense of the acquisition of associative relations acquired through trial-and-error learning, though not of the advanced cognition required for understanding an abstract concept like 'equivalence.'

Reasoning by exclusion can be carried out by way of a concept of not-same, but it can also be carried out non-conceptually. That is, reasoning by exclusion *need not* depend on an abstract concept (like equivalence) because it can be grounded solely in the particular stimuli itself. The organism's prior reinforcement history for the familiar choice stimuli has resulted in a trained association between that (incorrect) familiar choice stimulus and something other than the novel sample

stimulus because the principle behind trial and error says that when success is met to use that solution the next time the problem is encountered. The sample is not the known 'solution' to the familiar choice stimulus that has been arrived at previously, so the response to the novel sample stimulus is to select the novel choice stimulus—and at this stage in the game it does not even require trial and error because it is the only choice that lacks any reinforcement history. Associative learning (through previous trial and error) could explain the subject's correct response without reference to their possession of an abstract understanding of the relation of 'equivalence,' or, 'not-same.' Kastak and Schusterman suggest an easy way to protect against the subject's making use of exclusion in a MTS procedure: rather than presenting the comparison stimulus ♥ along with the familiar ♣ (or ♦), it should be presented with another novel stimulus (♠, for example).

Schusterman and colleagues' ongoing training and testing of a California sea lion suggests that exclusion can play a significant role in the subject's acquisition of an abstract understanding of the equivalence relation. Their sea lion subject utilized the exclusion effect in solving the initial identity problems she was confronted with.[45] But with experience an appreciation of the abstract relation of equivalence emerged that she was able to extend to contexts in which exclusion was no longer available to guide her choice of the correct comparison stimuli.

[45] As part of the training program, the sea lion subject was provided with the opportunity to utilize the exclusion effect when learning the relations of the first 8 of the 30 problem sets she was eventually tested on. For these first 8 sets of relations, a novel sample stimulus was always paired with a familiar comparison stimulus. This gave the subject the opportunity to grasp the tasks being asked of her on the later tests when exclusion was no longer available to guide her choice.

IX. Logical inference in a California sea lion.

Schusterman and Kastak (1993) are the first researchers to demonstrate the formation of equivalence relations between perceptually disparate visual stimuli in a nonhuman animal. In human subjects, training of one-way sign-referent relationships (aRb, bRc) produces emergent reflexive (aRa, bRb, cRc), symmetric (bRa, cRb), and transitive (aRc) relationships (Gisiner and Schusterman, 1992). That is, when a subject learns to associate *b* with *a* and *c* with *b*, they then spontaneously (i.e. without being explicitly taught) infer that the other relations hold as well. Sidman and Tailby (1982), and Gisiner and Schusterman (1992) following them, identify these three relationships as forming a set of stimulus equivalence relations between signs and referents. Utilizing the mathematical definition of an equivalence relation (reflexivity, symmetry, and transitivity) between two perceptually different stimuli, Sidman and Tailby have developed an expansion of the conditional-discrimination testing paradigm that allows researchers to determine whether an equivalence relation has emerged (without additional training or differential reinforcement) from the trained conditional relation.[46]

[46] Sidman (1971) originally conducted research on the emergence of equivalence relations from trained conditional relations on a severely mentally disabled human child who was unable to read either out loud or to himself with comprehension. The child was first taught to match pictures to spoken words (for example, a picture of a cat and the word 'cat') and was then taught to match spoken words to printed words (the spoken word 'cat' and the printed word 'CAT'). Subsequent to this training, the child was capable of matching the picture of a cat to the printed word 'CAT' without being explicitly trained to do so, demonstrating the acquisition of learned auditory-visual equivalences via transitive inference.

If a relation (aRb) is reflexive, then each stimulus bears the conditional relation to itself (aRa, bRb). Whether the relation of reflexivity has emerged can be tested with an identity matching procedure in which the subject is required to match stimuli *a* to itself and stimuli *b* to itself. If a conditional relation (aRb) is symmetric, the antecedent and consequent are reversible (bRa). If a subject matches sample *a* to comparison *b*, they must then (without training) go on to match sample *b* to comparison *a*. If two conditional relations (aRb, bRc) are transitive, then the consequent of one statement is the antecedent of the other. If this is the case, then the antecedent of the first statement indicates the consequent of the second statement (aRc). If a subject who has been taught the two conditional relations (aRb, bRc) is able to generate the conditional relation (aRc), the emergence of transitivity has been displayed.

Schusterman and Kastak have experimentally demonstrated the acquisition of the relations of reflexivity, symmetry, and transitivity in a nonhuman animal. Their subject was a captive adult female California sea lion, Rio. Rio was trained and tested on an artificial language that utilized visual stimuli in the form of black silhouettes of figures on a white background that comprised 30 different 3-member sets of perceptually disparate and arbitrarily combined symbols. For example, problem set number 16 was composed of the images of (*a*) a crab, (*b*) a tulip, and (*c*) a radio. The testing apparatus was a tri-partite board placed in front of the subject. The sample stimulus (i.e. the one Rio is being asked to match another stimulus to) appeared on the central section, directly in front of the subject. The two comparison

stimuli (i.e the options Rio was given to match to the sample stimulus, only one of which is the correct choice) appeared on the sections to either side of the sample stimulus. The two comparison stimuli were presented to the subject simultaneously while the sample stimulus remained exposed. After an interval of 2-4 seconds, the experimenter would release the subject from her station in front of the central section and the subject would point her nose at her choice.

The particular sequence of tests conducted were designed to maximize Rio's correct performance on test trials by ensuring that she had demonstrated all of the prerequisites for a given test before that test was given. After training of the aRb relation, the test for bRa symmetry could be given. After training the bRc relation, the test for cRb symmetry could be given. Following success at both the aRb and bRc relation, the test for aRc transitivity could be given. Only after acquisition of the aRc transitive relation could the test for cRa symmetry be given. After this series of tests had been passed by Rio for the first 12 of 30 problem sets (each composed of three members: (*a*), (*b*), and (*c*)), the prerequisites for the equivalence test had been demonstrated. The equivalence test was then conducted using the remaining problem sets 13-30. This sequence of training and testing ensured that the relation that Rio would acquire between the three perceptually disparate stimuli would be one of equivalence and that she would thereby be in a position to understand the task put to her in test trials. By first teaching her the conditional relations between stimuli (*a*) and (*b*) and between stimuli (*b*) and (*c*) followed by preliminary testing on her ability to spontaneously infer the symmetric and transitive consequences of those conditional

relations, the researchers demonstrated that Rio had acquired an equivalence concept that she was then able to transfer and apply to the remaining 18 problem sets.

Rio learned the first 8 aRb relations by exclusion, that is, when presented with a novel sample stimulus, the novel comparison stimulus was paired with a familiar comparison stimulus so Rio could use exclusion to select the correct comparison stimulus. For example, after learning problem set 1, which was composed of (a) a ring, (b) a baseball bat, and (c) a hanger, these stimuli were all familiar to her. So for problem set 2, which consisted of (a) a plus sign, (b) a square, and (c) a spatula, the aRb relation could be taught be presenting the a-member of problem set 2 (i.e. the plus sign, which was novel) and having her select her choice between the b-member of problem set 2 (the square, which was also novel) and a familiar stimuli from problem set 1 (the hanger, for example). The remaining 22 aRb relations were learned by trial and error; Rio would make a random choice and would be rewarded for selecting the comparison stimulus of the conditional relation on which she was being trained. She learned all 30 bRc relations by trial and error. Differential reinforcement was used in both cases; i.e. she was provided with a fish reward only for correct responses. Preliminary (i.e. prerequisite) testing on the relations of symmetry and transitivity were then performed on the 12 exemplar sets in order to assess Rio's ability to spontaneously infer the logical consequences of those trained conditional relations. This preliminary testing also served as Rio's only explicit training on those logical relations. On the tests for bRa symmetry for the first 12 aRb relations, Rio passed 8 of 12 problems (3 of 6 on the first test and 5 of 6 on the

second test). Although her overall performance was not significantly above chance, it should be acknowledged that she did show improvement on the second test. On the cRb symmetry tests for the first 12 bRc relations, Rio passed 10 of 12 problems (5 of 6 on each of two tests). On the test for transitivity for the first 12 aRb and bRc relations, Rio passed 11 of 12 problems (5 of 6 on the first test and 6 of 6 on the second test). On the cRa symmetry tests for the first 12 aRb, bRc, and aRc relations, Rio passed 10 of 12 problems (4 of 6 on the first test and 6 of 6 on the second test).

After Rio's successful performance on these preliminary tests for her ability to infer the logical consequences of conditional relations, "The final test conducted in the series assessed Rio's ability to combine transitive and symmetrical relation abilities in order to form equivalence relations on 18 potential 3-member stimulus classes without having had any previous symmetrical or transitive experience with them" (Schusterman and Kastak, 1993, 833). On this final cRa equivalence test, Rio passed 14 of 18 problems. These results indicate that, for Rio, an equivalence relation had formed between the 3 perceptually disparate stimuli ((*a*), (*b*), and (*c*)) of each stimulus set. In their meta-analysis of the literature on equivalence relations published in the time since Sidman's seminal 1971 paper on the topic, O'Donnell and Saunders (2003) provide the following evaluative standards for assessing a subject's test performance: the lower limit of the range of successful demonstration of the emergence of equivalence relations is calculated by adding 100 to chance and dividing by 2. A statistically significant finding on an equivalence test is thereby defined as at or above the midway point between chance and perfect performance. In

Rio's case, with two comparison stimuli, chance performance is 50, so any score above 75% would constitute success. Rio's score on the final tests for equivalence of 14 correct responses of 18 trials (77.78%) is therefore a successful demonstration of the emergence of equivalence relations on this method of assessment.

Sidman and Tailby (1982) have suggested that despite the apparent difficulty of nonhuman animals to form equivalence relations, by providing them with numerous examples we may facilitate the emergence of an understanding of symmetrical relations between perceptually disparate stimuli. Schusterman and Kastak state, "We believe that the critical factor in Rio's subsequent performances in passing tests of symmetry, transitivity, and equivalence stems directly from her experiencing enough exemplars to grasp these interrelated concepts. Thus, after being taught [on problem sets 1-12] that a number of samples and comparisons are interchangeable, Rio rapidly learned to respond to novel symmetrical relations [of problem sets 13-30] the first time she encountered them" (Schusterman and Kastak, 1993, 836).

X. Middle-Way: equivalence relations emerge from conditional relations on basis of exclusion

According to Samuels and Stich, the body of experimental data on human subjects' abilities and failures to solve reasoning tasks indicates human cognition is supported by two distinct neuronal systems, one of which is constituted by a number of domain-specific mental modules together with a small set of stimulus-bound rough

heuristics and another system that is constituted by a set of abstract, domain-general problem-solving skills. As with the experimental data on human reasoning and decision making, it is my position that the middle-way is also better able to account for the growing body of marine mammal data. The old path, being fast and unconscious, accounts for the sea lion's initial use of trial and error and exclusion when confronted with a conditional discrimination task. The new path, being slow, conscious, and highly influenced by training and education, accounts for the emergence of equivalence relations from those conditional relations and, thereby, for the sea lion's capacity for logical inference.

The middle-way's reference to dual-process theories sheds light on the mechanisms of mind involved in the sea lion subject's eventual acquisition of an equivalence concept. The evolutionarily older system is responsible for our ability to solve concrete problems in our environment using simple content-independent rules of thumb, or rough heuristics (e.g. trial and error), as well as context-dependent, domain-specific reasoning procedures (e.g the evolutionary psychologists' Cheater Detection Module[47]). The newer system, which is frequently presumed to be uniquely human, is largely under conscious control and is heavily influenced by training in its application. This latter system is proposed to be responsible for our ability to reason in accordance with the abstract inferential rules of the formal theories of deductive logic, probability calculus, and decision theory.

[47] Leda Cosmides, "The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task," *Cognition*, vol. 31, no. 3, 1989, pp. 187-276.

The idea that the faster, unconscious system evolved earlier than the slower, controlled system predominates throughout the dual-processing literature, as does the claim that although the older system is shared with other animals, the newer system is responsible for the 'uniquely human' characteristics of higher-order thought (i.e. reflective consciousness, abstraction, linguistic competence). But evidence indicates that there is a distinction between stimulus-bound and higher-order processing in many species, including primates and rodents (Toates, 2006) in addition to marine mammals (Schusterman and Kastak, 1993). While Rio initially relied on trial and error and later on inference by exclusion in her selection of the correct comparison stimuli, after a sufficient amount of practice on the first 12 problem sets, the concept of equivalence emerged and she was able to automatically apply what she had learned from that training to the remaining 18 problem sets on which she had no direct training (i.e. reinforcement history) on the relations of symmetry and transitivity. Samuels and Stich state,

> Since the fast, automatic, evolutionary older system requires little cognitive capacity, everyone has the capacity to deal rationally with many reasoning and decision making problems that were important in the environment in which we evolved. Moreover, since the new, slow, rule based system can be significantly affected by education, there is reason to hope that better educational strategies will improve people's performance on those problems that the old system was not designed to deal with. (Samuels and Stich, 298).

Although the older system is presumed to lack the plasticity of the newer system, with proper training we can learn to inhibit the older system. That is, via education

and experience our reliance on the older system is replaced by the advanced cognitive abilities that emerge with the development and refinement of the newer system.

The stimuli Rio's artificial language is composed of has no intrinsic biological significance to sea lions. But the training procedure ensured that she had met the prerequisites for competence before a test was given. Equivalence relations emerged from training, despite context and content remaining foreign to a sea lion's natural environment and thereaby unable to be accounted for by stimulus-bound behaviors that are either reflexive or acquired via conditioning. As with college students in their first class on formal logic, training and practice is essential to master these types of problems that our evolutionary history has not had the requisite time to equip us with specialized tools to solve. But those evoluionarily ancient problem-solving strategies we do possess can be co-opted and refined with the proper education.[48] In the case of Rio, unlike the evolutionary psychologists' subjects, even though the content was not altered to mimic a realistic problem sea lions face in their natural environment, she was able to perform successfully on the novel reasoning tasks she was presented with, once she'd received enough exemplar training.

[48] As Colin Allen points out, "Transitive relationships are frequently important to animals, especially those living in social groups" (Allen, 175). Social dominance relations, for example, "provide a domain in which a capacity for transitive inference would seem to be very useful" (Allen, 176). Social knowledge, or knowledge of the social dynamic of the group, can be acquired by an individual not only through their own interactions with others, but also through their observations of others interacting. If an individual *a* has been dominated by a conspecific *b* and observes conspecific *c* dominate conspecific *b*, then they may judge that conspecific *c* is dominant over them based on the relation of transitivity. As Allen notes, "Unlike dominance hierarchies, the experimenter-imposed ordering on stimuli has no biological significance to the animals nor any connection to any naturally occurring transitive relationship" (Allen, 181).

XI. Conclusion

Rio's ability to acquire an equivalence concept in a biologically insignificant context indicates that, in sea lions, such inferential reasoning is not restricted to a specific domain. Perhaps it is a by-product of a domain-specialized mechanism, but it is not domain-restricted in that it can be co-opted and utilized in an alternative context when sufficient and appropriate training is provided. Although Fodor (1983) introduced the concept of modularity of mind, he did so in the context of a dual-processing theory that distinguished the input of particular capacities like vision and language from the mechanisms underwriting general-purpose, central cognition. Like Samuels and Stich, Fodor (2001) has attacked the 'massive modularity hypothesis' that has been proposed by evolutionary psychologists, arguing that the claim that all (or at least most) reasoning and decision making in human subjects is the result of domain-specific mental mechanisms is not supported by the data. Rather, Fodor has proposed that we rely on both general-purpose and specialized mental mechanisms in many reasoning and problem-solving tasks. Data coming from Schusterman and colleagues' studies of logical inference in California sea lions suggests that the same is true of these marine mammals.

Massive modularity of mind is not a characteristic of the human mind, so it cannot be what distinguishes humans from all other animals. The middle-way's portrayal of human rationality as arising from both domain-general cognitive capacities and domain-specific mental modules best accounts for the data on logical inference in sea lions as well. Although dual-processing theorists frequently presume

that some nonhuman animals do possess lower-level cognitive processes, they propose that what distinguishes humans from all other animals is the development of the higher-level system. But as evidenced by the work done with Rio, when appropriate elicitation of the higher-system and suppression of the lower-system is provided, some nonhuman animals have been shown to perform tasks that rely on advanced cognitive processing.

XII. Implications and future directions

Animal behavior and cognition research could benefit from a shift of focus from the demonstration of species-typical cognitive traits to an investigation of the potentials of nonhuman animal minds. Despite the fact that Rio's capacity to perform ecologically irrelevant reasoning tasks is very much the product of the specific training and testing procedure she has been immersed in and may not be expressed and functioning as a species-typical trait of California sea lions, her demonstration of complex cognition shows that sea lions possess the requisite neurophysiological structures to allow the elicitation and emergence of those potentials in the form of expressed competencies. By investigating the permeability of domains of reasoning rather than simply attempting to demonstrate species-typical traits (resulting from evolutionary psychology's attempt to delineate the limits of their presumed domains), animal behavior and cognition researchers may illuminate important aspects of the structure of both human and nonhuman animals' minds.

Chapter 4: Associationism and Primate Theory of Mind

I. Associationism (revisited)

Even the staunchest supporter of associative accounts of human cognition agree that certain purportedly uniquely human capacities, formal reasoning for example, require non-associative mechanisms to be accounted for. So the work done with Rio on a California sea lion's capacity to perform logical reasoning tasks indicates that some nonhuman animal species can (and, under experimental conditions, do) make use of higher-order cognitive mechanisms. Although much of what appears to require explanations that reference higher-order processing may in fact be carried out by simpler, associative mechanisms, there is a subset of those abilities that do in fact require us to postulate advanced cognitive processes, both in humans and in other animals.

As discussed in Chapter 1, with his Canon, C.L. Morgan did not intend to forbid reference to higher-order cognitive processes when explaining the apparently complex nonhuman animal behavior. Rather, he meant to restrict such references to those cases where, together with our greater body of knowledge, we are unable to account for some behavior/performance by postulating only simpler cognitive mechanisms.[49] It is therefore worth reexamining the claims made by the dissenters

[49] Morgan's intention has been lost in the centuries-long adherence in ABC research to the Canon. The revised Canon added a clause about our entire body of knowledge. So seeing that when utilizing an ecologically-invalid paradigm evidence of abstract reasoning in a non-ape species was demonstrated, we should reexamine the primate ToM findings in light of this. When assessing the entire body of experimental data we have on nonhuman animals' ability to engage in nonassociative cognitive processing, it becomes evident that a purely associationist account of the findings cannot be justified by the data.

in the animal minds debate that the primate theory of mind findings can be accounted for by purely associative mechanisms.

Before we can take a deeper, critical look at the primate ToM literature, we must first get clear on a precise account of what associationism about some cognitive feat entails. That is, we require a clearly articulated definition of associationism as well as an explicit method for determining whether some skill/behavior is in fact underwritten by such a mechanism. After laying out the philosophical and psychological history of the concept of 'associationism,' Eric Mandelbaum's explication of the various forms of associationist theses will be reviewed and brought to bear on each of the primate ToM experiments discussed in Chapter 2. I aim to show that once we understand associationism in its distinct forms, it becomes difficult to make sense of some of the primate theory of mind experiments in an associationist framework.

II. Associationism

Broadly speaking, 'associationism' refers to psychological theories that attempt to explain apparently complex cognition as being built-up from simple associations between sensations/stimuli and behavior/responses. In philosophy, the emphasis has been primarily on the transition between thoughts (i.e. association of ideas) whereas in psychology it has been on how sensations give rise to behavior (i.e. association of responses to stimuli).

i. Associationism in philosophy (Empiricism)

Empiricists offered associationism as a theory of both learning and thinking, arguing that all concepts are acquired through experience, but the mental processes that underwrite such learning were rarely posited to have been learned. Although Locke's discussion of the 'association of ideas' in his fourth edition of the *Essay Concerning Human Understanding* (1700) inaugurated discussions on the topic, it was Hume that provided the first detailed account of associationism as a theory of learning in his *Treatise of Human Nature* (1739). He states, "[A]ll our simple ideas in their first appearance are deriv'd from simple impressions, which are correspondent to them, and which they exactly represent" (Hume, *Treatise of Human* Nature, 1739, Book 1, Part 1, Section 1, par. 7). In this text Hume aimed to lay out how perceptions (i.e. "Impressions") determined trains of thought (i.e. successions of "Ideas"). According to Hume, "it be too obvious to escape observation, that different ideas are connected together; I do not find that any philosopher has attempted to enumerate or class all the principles of association" (*Enquiry Concerning Human Understanding*, 1748, 24). Hume introduced the notion of associative learning, which is still prevalent in psychological and philosophical theorizing today.

According to Hume's Copy Principle, "there were no Ideas in the mind that were not first given in experience," …"the ordering of Ideas was determined by the ordering of the Impressions that caused the Ideas to arise" (Mandelbaum, 'Associationist Theories of Thought,' SEP, 2015, sec. 3, par. 1). Hume's Copy Principle succinctly encapsulates the basic premise of empiricism; it accounts for the

origin of ideas by saying that all Ideas are copied from Impressions we receive in experience. Simple ideas can combine into complex ideas, but the impressions from which these simple ideas arise are acquired in experience. It is the mechanism of association that accounts for the connection of such ideas.

For Hume (1739), associationism was a theory of mental processes according to which the only mental process is the ability to associate ideas, and this ability is able to account for both learning and thinking. Hume posited three principles of association to account for the observation that different ideas are connected to one another: cause and effect, contiguity (i.e. proximity), and resemblance. These principles are neither rational nor theoretical, but are understood as natural operations of the mind. Uniting the Copy Principle with the three principles of association, Hume proposed that two Ideas in the mind would hold the same relation that was instantiated by the two Impressions that gave rise to them. Given his disdain for metaphysics, the three principles of association are to be taken as primary "original qualities of human nature" (*Treatise*, 13), and trying to account for them takes one beyond the bounds of experience.[50]

Hume further distinguished 'relations of ideas' from 'matters of fact,' the former of which includes mathematical and logical reasoning and the latter refers to perceptual inferences about causal relations. Although both 'relations of ideas' and

[50] Hume embraces empiricism's central tenet that all knowledge is acquired through experience, even our knowledge of human nature. It follows that, "The essence of the mind being equally unknown to us with that of external bodies, it must be equally impossible to form any notion of its powers and qualities otherwise than from careful and exact experiments, and the observation of particular effects, which result from different circumstances and situations" (Hume, *Treatise*, xvi). With this, Hume introduced an experimental science of human nature that established the psychological sciences as we know them today.

'matters of fact' are, according to Hume, based in associations, nonhuman animals engage in only the latter whereas humans engage in both forms of reasoning. He further proposed that two objects that are not immediately related in experience might come to be associated in the imagination. He states, "two objects are connected together in the imagination, not only when the one is immediately resembling, contiguous, or the cause of the other, but also when there is interposed betwixt them a third object, which bears to both of them any of these relations" (Hume, *A Treatise of Human Nature*, 1739, Book 1, Part 1, section 4). It is Hume's Copy Principle together with his principles of association that are distinctive to his empiricism.

ii. Associationism in psychology (conditioning)

Ivan Pavlov modernized Hume's account of associative learning with his concept of classical conditioning[51], which initiated the behaviorist movement in modern psychology (both human and nonhuman animal). Classical conditioning is a general method of learning in which an unconditioned stimulus (US)[52] is paired with a novel neutral stimulus. With repeated exposures, the contiguity of the US and the neutral stimulus causes the latter to provoke the same response as the former, thereby becoming a conditioned stimulus (CS). Classical conditioning is a stimulus substitution paradigm by which the CS gets associated with the response to the US, but the response itself remains unchanged. In Pavlov's canonical experiment, the US

[51] Conditioning is the process of acquiring new associations.
[52] An US (unconditioned stimulus) is a stimulus that instinctively, without training, provokes a response in the organism, i.e. the UR (unconditioned response).

was meat powder, the unconditioned response (UR) was salivation, and the CS was a bell. With repeated trials, the CS of the bell gave rise to the conditioned response (CR) of salivation in his canine subjects.

It was Edward Thorndike's work on instrumental conditioning with cats in puzzle boxes that expanded the theory of associative learning beyond instinctual behaviors and sensory substitution by introducing the notion of consequences, allowing him to produce novel behaviors in his animal subjects. Thorndike's cats had to learn responses that were not instinctual behaviors (unlike Pavlov's URs) in order to escape the "puzzle boxes"[53] they were trapped in, and the behaviors that they acquired were shaped by the consequences that resulted from such behaviors. If the act of pressing a lever caused the door on the puzzle box to open, then the cats would learn the connection between the lever and the door. Whereas classical conditioning is a learning paradigm in which stimuli are associated with other stimuli, instrumental conditioning is a learning paradigm in which stimuli are associated with responses. The latter provides a method for eliciting novel behaviors in human and nonhuman animal subjects.

In contrast to the passive learning that occurs in a classical conditioning paradigm, Thorndike's instrumental conditioning paradigm (as well as the operant

[53] Thorndike's puzzle boxes were approximately 20 inches long, 15 inches wide, and 12 inches tall and were completely enclosed. They were each equipped with a door that could be pulled open by pressing a lever or pushing a button inside the box, which caused a weight attached to a string that ran over a pulley to open the door. Thorndike found that his animal subjects would, with time, arrive at the solution to the puzzle box by chance/accident, but that then on subsequent trials on the same problem they would perform the required behavior more deliberately and with less of a time delay after being placed in the box. He posited that they relied solely on learning by trial-and error to arrive at the correct solution to the problem they were faced with.

conditioning paradigm that developed from it) is a species-nonspecific, general, and active theory of learning because it does not depend on a species' instinctual response to a stimulus (Mandelbaum, SEP, 'Assoc.,' 2015, sec. 3, par. 6). Thorndike's "Law of Effect" (1911) was the first canonical psychological law of associationist learning. It proposed that,

> [R]esponses that are accompanied by the organism feeling satisfied will, *ceteris paribus*, be more likely to be associated with the situation in which the behavior was executed, whereas responses that are accompanied by a feeling of discomfort to the animal will, *ceteris paribus*, make the response less likely to occur when the organism encounters the same situation. (Mandelbaum, SEP, 'Associationist Theories of Thought,' 2015, s 3, p 7)

For Thorndike, the process of acquiring associations (i.e. conditioning) is analogous to the process of natural selection, making it akin to non-cognitive reflexes, except that it occurs within an individual organism's life history. Thorndike's "Law of Exercise" states that, "responses to situations will, *ceteris paribus*, be more connected to those situations in proportion to the frequency of past pairings between situation and response" (Mandelbaum, SEP, 'Assoc.' sec. 3, par. 7). Taken together, Thorndike's laws of associationist learning represent the modern incarnation of empiricist thought as it applies to nonhuman animal behavior.

Following Watson's proposal to make the study of psychology scientific by utilizing only objective procedures (i.e. measuring behaviors), B.F. Skinner (*The Behavior of* Organisms, 1938 and *Science and Human* Behavior, New York: Macmillan, 1953) introduced the term 'operant' conditioning to replace the instrumental conditioning of his predecessors. Skinner's operant conditioning

paradigm supplemented Thorndike's notion of consequences with that of 'reinforcement' and 'punishment', and in doing so eliminated any reference to even simple mental states such as 'satisfaction' and 'discomfort.'[54] According to Skinner,

> [M]y research was on the role of the consequences of behavior, and that was 'learning.' They were not, however, the consequences that lay ahead in a particular instance as the goal or purpose of the behavior; they were the consequences that had followed behavior in the past. (B.F. Skinner, "From Behaviorism to Teaching Machines to Enjoying Old Age," *This Week's Citation Classic*, number 6, Feb. 5, 1990, par. 4)

Whereas Thorndike's experiments had led him to the notion of trial-and-error learning, Skinner proposed that in his experimental work "the organism was not necessarily trying to do anything, and it certainly learned by successes rather than failures or errors" (Skinner, 1990, par. 5). With operant conditioning, a response to a given stimulus is modified by prior reinforcement and punishment, which are defined by their effects on behavior; reinforcement increases a behavior and punishment decreases a behavior. In both cases, the strength of a behavior is modified by its consequences. Extinction is the only other method by which behavior is modified on an associative account of learning.

Positive reinforcement occurs when a behavior/response is itself rewarding or is followed by another stimulus that is rewarding. It increases the frequency of a behavior and is often referred to simply as 'reinforcement.' An example is a rat learning to press a lever to get a food reward. Negative reinforcement occurs when a behavior/response is followed by the removal of an aversive stimulus. It increases the

[54] Skinner's anti-representationalism marks a departure from earlier associative theorizing and best demonstrates the difference between associationism and behaviorism.

frequency of a behavior and is often referred to as 'escape.' An example is a rat learning to press a lever to turn off an uncomfortably loud noise. Both positive reinforcement (reinforcement) and negative reinforcement (escape) increase the frequency of the behavior performed by the animal in the given context in which learning occurred.

Positive punishment occurs when a behavior/response is followed by an aversive stimulus. It decreases the frequency of a behavior and is often referred to simply as 'punishment.' An example is touching a hot stove and being burned by it, or performing some undesired behavior and receiving a spanking. Negative punishment occurs when a behavior/response is followed by the removal of a rewarding stimulus. It results in a decrease in the frequency of the behavior and is often referred to as 'penalty.' An example is taking a favorite toy from a child for their undesired behavior. Both positive punishment (punishment) and negative punishment (penalty) decrease the frequency of the behavior in the context in which the learning took place.

Immediacy, sometimes referred to as 'contiguity' by psychologists, refers to the temporal proximity of the reinforcement or punishment to the behavior that elicits it. The more immediate a consequence, the more effective it will be in modifying the behavior.[55] Contingency refers to the consistency of the consequences for a given

[55] Since Hume, contiguity has been central to associationist accounts of thought. The problem of determining the parameters needed for contiguity has been referred to as the problem of the "Window of Association." For contiguity to serve as a founding pillar of associationism, then the window needs to be relatively short. Additionally, "if the domain generality of associative learning is desired, then the window needs to by homogenous across content domains" (Mandelbaum, "Associationist Theories of Thought," SEP, 2015, 9.4, par. 1).

behavior.[56] Reinforcement or punishment that occurs consistently after the behavior/response and not at other times is most effective. If the reinforcement had been intermittent, extinction of the response takes significantly longer.

Extinction occurs when a behavior/response that had previously been reinforced is no longer reinforced and with repeated trials the frequency of the behavior decreases. It differs from punishment because neither an aversive stimulus is being applied nor is a rewarding stimulus being withheld. For example, a rat who had been given food many times for pressing a lever so that the frequency of the behavior had increased as a result of conditioning via positive reinforcement is then no longer given the food reward for pressing the lever. The rat will press the lever more and more infrequently until the behavior stops altogether, at which point the lever pressing (i.e. conditioned response) has been "extinguished." It is important to keep in mind that operant conditioning does not only occur in the laboratory. Naturally occurring consequences, or the lack thereof, can reinforce, punish, or extinguish behaviors.

iii. Associationism and behaviorism

According to J.B. Watson, the purpose of psychology is, "To predict, given the stimulus, what reaction will take place; or, given the reaction, state what the situation or stimulus is that caused the reaction" (John B. Watson, *Behaviorism*, New

[56] In the Rescorla Wagner (1972) model, the contingency requirement supersedes both contiguity and resemblance. If 'resemblance' refers to the resemblance amongst contents of thought, as it did for Hume, then it makes sense that this requirement would be abandoned by those arguing for associative learning explaining apparently complex behavior in the absence of conscious mental representation, i.e. thought.

York, Norton, 1930, p. 11). The radical behaviorism of Skinner is committed to three claims, each of which can, in principle, be taken on its own without adherence to the others. According to the *methodological* tenet, psychology is and should be the science of observable behavior. According to the *psychological* tenet, behavior can be explained without reference to mental events or internal psychological processes. And the *analytical* tenet proposes that mental terms or concepts should be eliminated and replaced with behavioral terms, or should be translated into behavioral concepts. (Graham, SEP, 'Behaviorism,' 2016, section 1. Par. 5)

In contrast, associationism as a general theory of cognition is neutral on whether associations are implemented at the representational (i.e. psychological) level or at the physical (i.e. neural) level (Mandelbaum, SEP, 'Assoc.,' 2016, section 6, par. 1). But the psychological behaviorism that arose from the empiricists' associationist account of cognition explains human and nonhuman animal behavior solely in terms of external physical stimuli, responses, learning histories, and reinforcements. For the behaviorist, reference to the role of mental representations is not only unwarranted, but is also unnecessary when explaining the apparently complex behaviors of humans and other animals. According to Graham,

> Classical [philosophical] associationism relied on introspectible entities, such as perceptual experiences or stimulations as the first links in associations, and thoughts or ideas as the second links. Psychological behaviorism, motivated by experimental interests, claims that to understand the origins of behavior, reference to stimulations (experiences) should be replaced by reference to stimuli (physical events in the environment), and that reference to thoughts or ideas should be eliminated or displaced in favor of reference to responses (overt behavior, motor movement). Psychological

> behaviorism is associationism without reference to mental events" (Graham, SEP, 'Behaviorism,' 2016, section 3, par. 4).

Behaviorism is distinct from, but deeply connected to, associationist theorizing. But, despite the important differences, in animal behavior and cognition research the two remain irremediably intertwined. As we will see when reexamining the primate ToM experiments, for both the proponents and the naysayers in the animal minds debate, 'associationism' is interpreted as 'behaviorism.' This is likely a result of the field of animal behavior and cognition research's strict adherence to a particular interpretation of Morgan's Canon, one which bars reference to mental events if a particular performance can be explained by reference only to physiological and/or simpler cognitive processes (i.e. innate reflexes and/or conditioned responses). It is an unfortunate consequence of the conflation of 'associationism' and 'behaviorism' that when the claim is made that some performance can be explained associatively, without reference to higher-order cognitive processes, what is actually being claimed is that no reference to the internal workings of the animal subject's mind is required for a successful and thorough account of the cognitive mechanisms underpinning the given behavior. But Morgan himself was not a behaviorist in either the radical sense or even in the methodological, psychological, or analytical sense. Given Morgan's own willingness to posit internal mental processes when doing so is required to make sense of our entirety of knowledge on a subject, those working under his namesake methodological principle should examine the larger body of data when interpreting the results of a given experiment, rather than attempting to explain each experimental result in isolation.

III. Mandelbaum on the varieties of associationist theses

It is remarkable how varied the use of the term "association" is across the different divisions of the university: from the Humanities to the Social Sciences to the Biological Sciences, and so on. Because of this variety of use, it can be difficult to determine whether the term is being used to refer to the same phenomenon, to different phenomena, or to various aspects of the same phenomenon. But just because a usage has become common does not entail that it is not a misappropriation of a scientific term with a specific and definable meaning. It would be a benefit for all engaged in the nonhuman animal minds debate to arrive at a common definition of the term 'associationism' and to disambiguate it from 'behaviorism.'

Eric Mandelbaum has provided a useful discussion of the varieties of associationist theses (and the conflation of them) that have appeared in the human social psychological literature, specifically regarding the mechanisms underlying implicit biases. In addition to distinguishing a variety of types of associationist theses, Mandelbaum also provides a method for determining whether or not some behavior is in fact underwritten by associative mechanisms; he does this by seeing how purported associations change, or do not change, under certain conditions. Although Mandelbaum's focus within the cognitive and behavioral sciences differs from my own, his work can be extended to the associationist arguments put forward in the nonhuman animal minds debate.

According to Mandelbaum, 'associationism' refers to a theory of thought, but not a single theory of cognition,

> [R]ather a constellation of related though separable theses […] that share a commitment to a certain arationality of thought; a creature's mental states are associated because of some facts about its causal history, and having these mental states associated entails that bringing one of a pair of associates to mind will, *ceteris paribus*, ensure that the other also becomes activated. (Mandelbaum, SEP, 'Associationist Theories of Thought,' 2015, par. 1)

An association is a relation between ideas (on the empiricist account) or between sensations and responses (on the behaviorist account). What all adherents to an associationist account share is the belief that an organism's experience is the main sculptor of their cognitive architecture.

In contrast to the faculty psychologist or modularity theorist, who would claim that the question of how many mental processes there are cannot be explained *a priori*, the pure associationist proposes that there is only one type of mental process and that is association whether of ideas or of stimuli and response (the latter of which does not depend on mental representation). Association is purported to be able to account for learning, the structure of mental states, and the way certain thoughts relate to other thoughts (i.e. associative transitions, which do depend on mental representation, even if not conceptual understanding necessarily). But, according to Mandelbaum, "the inference from one sense of association to another is invalid without further argument and evidence" (Mandelbaum, "Attitude, Inference, Association: On the Propositional Structure of Implicit Bias," *Nous*, 2014, p. 5). By positing a single mental process underlying thinking, learning, and cognitive

structure, the associationist purports to have parsimony on their side (at least, in a certain sense of parsimony). But, as Mandelbaum points out, accepting an associative account of one of these mental processes does not entail that we must also accept associative accounts of the others.

i. Associative learning/conditioning

Associationism can be utilized as a theory of learning, a theory of thinking, and/or a theory of mental structures. Classical and operant conditioning paradigms rely on associative learning. According to Mandelbaum, "what all varieties [of associative learning] share with their historical predecessors is that associative learning is supposed to mirror the contingencies in the world *without adding additional structure to them*" (Mandelbaum, SEP, 'Assoc.,' 2015, 3, par. 8; italics added). Mandelbaum cites the prevalence of claims for the domain generality of associative learning as due in part to the adherence to traditional empiricist notions because it limits the amount of innate mental processes one has to posit. But as Mandelbaum states, "Merely having behavior reinforced in traditional ways does not ensure that any associative structure will be acquired" (Mandelbaum, 2014, 6). That is, associative learning does not necessarily eventuate in an associative structure.[57]

[57] For example, as discussed in the previous chapter, the sea lion subject Rio learned 22 of the aRb relations by trial and error and she learned all 30 bRc relations by trial and error. Differential reinforcement was used in both cases; i.e. she was provided with a fish reward only for correct responses. So the majority of the training she received was via associative learning. But when preliminary testing on her ability to spontaneously infer the logical consequences of symmetry and transitivity of those trained conditional relations was performed for the first 12 stimuli sets, Rio passed 8 of 12 problems on the bRa symmetry test, 10 of 12 problems on the cRb symmetry tests, 11 of 12 problems on the test for transitivity, 10 of 12 problems on the cRa symmetry tests, and on the final cRa

ii. Associative structures/mechanisms

Associative learning describes what is associated, but the question of how such an association is stored is an additional question. Associationists posit the notion of an 'associative structure' to account for this. An associative structure is a type of relation that may hold between two distinct mental states or a mental state and a valence; "saying that two concepts are associated amounts to saying that there is a reliable, psychologically basic causal relation that holds between them—the activation of one of the concepts causes the activation of the other" (Mandelbaum, SEP, 'Assoc.' 2015, 4, par. 2). A purely associative relation between mental representations adds no additional structure to the contents of those representations. It is in this sense that an association is, in Jerry Fodor's (2003) terms, 'semantically transparent.' This is in contrast to propositionally structured relations between mental representations, which have hidden structure above and beyond the causal relation between the representations.[58]

---

equivalence test on the 18 remaining stimuli sets Rio passed 14 of 18 problems. These results indicate that, for Rio, an equivalence relation had formed between the 3 perceptually disparate stimuli ((*a*), (*b*), and (*c*)) of each stimulus set despite only having received explicit training on the aRb and bRc conditional relations. It follows that what Rio learned is not associatively structured because those trained associations were able to enter into untrained inferential processes, despite being acquired via an operant conditioning procedure.

[58] A single proposition can be expressed by numerous different statements (i.e. representations) and the same statement can express a proposition at some times and a simple exclamation at others. For example, both "2 + 2 = 4" and "Four is the sum of two and two" express the same proposition. It is the proposition, not the uttered statements, that has a truth-value because the proposition is the shared contents of those synonymous statements. Additionally, "Rats!" can be used as an expression of disappointment (which has no truth-value) or as a descriptive sentence with the contents that there are rats present (which does have a truth-value). It is the hidden syntactic structure (i.e. deep grammar) that allows for this in propositionally structured relations, but not in associatively structured relations.

Mandelbaum notes that the prevalent view, which takes associative learning to eventuate in associative structures, is not necessary; "Logically speaking, there is no reason to bar any type of structure to arise from a particular type of learning" (SEP, 'Assoc.,' 2015, 4.3, par. 2). According to Mandelbaum, "Associative structures can be doubly dissociated from associative learning: one can gain associative structures from non-associative learning and associative learning can directly lead to the acquisition of propositional structures" (Mandelbaum, 2014, 5). Mitchell et al. (2009) have discussed propositional structures arising from associative learning. They propose two ways this could happen: "one may gain an associative structure that has a proposition as one of its associates"; or "one might also just have a propositional structure result from associative learning" (Mandelbaum, SEP, 'Assoc.,' 4.3, 2). Propositions can be acquired through associative learning and can enter into associative transitions, but the propositions themselves (i.e. the things learned) are not associatively structured. But once learned, these propositions can also enter into logical/inferential transitions even though they were acquired via associative learning mechanisms.

iii. Associative transitions/thinking

A pure associationist needs to account not just for learning and cognitive structure, but also for the transition between thoughts. As described by Mandelbaum,

> Associative transitions are movements between thoughts that are not predicated on a prior logical relationship between elements of the thoughts that one connects. In this sense, associative transitions are contrasted with computational transitions as analyzed by the

> Computational Theory of Mind. CTM understands inferences as truth preserving movements in thoughts that are underwritten by formal/syntactic properties of thoughts. (…) Associative transitions are transitions in thought that are not based on the logico-syntactic properties of thought. Rather, they are transitions in thought that occur based on the associative relations among the separate thoughts. (SEP, 2015, sec. 5, par. 1)

That is, the associative relations among thoughts are acquired directly via experience, not some form of reasoning or inference.

Mandelbaum discusses the idea of an "associative inference," which he claims is a borderline oxymoron. He says that we can give sense to the idea of an associative inference by positing transitions in thought that began because they were purely inferential but that became associated over time. An example familiar to all undergraduate philosophy instructors is the transition from the premises 'Socrates is a man' and 'All men are mortal' to the conclusion that 'Socrates is mortal.' We no longer think of the logical relations between the three statements, i.e. make an inference, but rather the conclusion automatically springs to mind with the recitation of the premises. Indeed, the second premise itself is associated with the first premise.

IV. Applying Mandelbaum's treatment to the animal minds debate

In contrast to Mandelbaum's treatment of the implicit bias literature, in which he focuses on transitions and structures, it is associative learning that is generally the focus in debates over the mindedness of nonhuman animals. Claims for a behavior's basis in an associative (i.e. lower-order) cognitive mechanism is assumed to follow directly from the argument that it was acquired via associative learning. And

references to associative transitions are absent from such discussions. This is likely due to the fact that many who propose that all nonhuman animal cognition is associative have already ruled out the possession of conscious thought in nonhuman animals, or at least any reference to inner states when explaining their behavior. They seek to explain all animal behavior non-consciously.[59] Ultimately, they are arguing from a behavior's basis in associative learning to the claim that whatever was learned is necessarily associatively structured. This is taken to provide evidence against advanced cognition in nonhuman animals as well as to eliminate any need to mention the role of mental representations in explaining apparently complex behavior. But as Mandelbaum's exposition of the varieties of associative theses has demonstrated, the fact that some ability was acquired via associative learning (i.e. conditioning) does not necessarily entail that the thing learned is associatively structured and that it does not enter into computational transitions.

If 'behaviorism' is understood as associationism without reference to mental states, then behaviorism is a better description of associationism in regards to the positions put forward in the animal minds debate and in ABC research in general. Just as associationism is not a single theory, but a set of (3) related though separable theses, so too is behaviorism. The radical behaviorism espoused by Skinner and

[59] A distinction must be made between something that is 'unconscious' and something that is 'nonconscious.' 'Unconscious' refers to something that an organism that can be conscious of things is not conscious of. 'Nonconscious' refers to an organism that is not conscious of anything or to those things that we could not possibly gain conscious awareness of, like the physiological processes that are controlled by our nervous systems. Further, talk of consciousness is not particularly helpful when making claims about the associational nature of nonhuman animal cognition. This is because one can have conscious awareness of associations they harbor as well as having unconscious non-associatively structured contents (for example, propositions in the language of thought).

Watson incorporates a methodological principle, a psychological principle, and an analytical principle. Although an individual researcher may adhere to and refrain from any combination of these tenets, they are most frequently conflated in the primate ToM debate.

Behaviorist tenets have been largely abandoned in many fields that investigate nonhuman animal cognition (cognitive ethology, comparative psychology, ecological psychology, cognitive science, neuroscience). And in studies of human cognition, the role of representation between the environment and behavior is accepted as essential. Nevertheless, both methodological and psychological behaviorist strictures remain strong in laboratory-based animal learning theory. And due to the (often unjustified) move made from a behavior's basis in associative learning to the claim that that behavior is underwritten by associative mechanisms, those behaviorist tenets are reified in the analysis of experimental results.

i. Mandelbaum's method for determining whether a behavior is carried out by an associative mechanism

Knowing that some content was acquired via associative learning, we cannot infer that it is therefore associatively structured. We need a method for determining whether what we are dealing with is an associative structure. As stated by Mandelbaum,

> Though associative learning, transitions and structures can be dissociated from one another, there is a connection between them that will be of much probative value: how to modify associations. We can infer whether a given cognitive structure is associative by seeing how

> certain types of information modify (or fail to modify) behaviors under the control of the cognitive structures. (Mandelbaum, 2014, 6)

Mandelbaum has provided a method for determining whether a relation is associative, that is, whether a behavior can be explained by associative mechanisms.

Since Pavlov, associationist theorists have been clear on how to modulate an established association, via extinction or counterconditioning. In the case of a relation that was acquired through classical conditioning, extinction decouples the association between a CS and US by presenting the CS without the US, and sometimes the US without the CS. For example, to extinguish the relation between the bell and salivation in Pavlov's dogs we would need to repeatedly present the meat powder in the absence of the bell and/or the bell in the absence of the meat powder. With repeated exposure of the CS in absence of the US, or vice versa, the animal will learn to disconnect the former from the latter. In the case of a relation acquired through operant conditioning, extinction breaks the connection between the operant response and the reinforcement or punishment by withholding the expected/conditioned result. In the case of an animal that had been conditioned to press a button to terminate an aversive stimulus, we would need to repeatedly refrain from eliminating the aversive stimuli after the button had been pushed.

Associative learning can condition stimulus/response associations and stimulus/stimulus associations. Mandelbaum states, "Just as we'd destroy stimulus/response associations through changing certain external contingencies, so too can we change stimulus/stimulus associations, the co-occurrence of certain representations, through changes in the external stimuli" (Mandelbaum, 2014, 6).

Extinction is one of two methods of doing this. If an association has been formed via positive reinforcement, then we can refrain from rewarding the trained behavior and in time it will be eliminated. That is, you can extinguish an associative structure by presenting one of the relata without the other.

The other method of breaking apart an associative structure is counterconditioning. Counterconditioning involves changing the valence of the relata of the associative structure. For example, if an animal subject receives a food reward for pushing a lever, then we can counter condition that association by providing a shock instead of the reward. Counterconditioning is a form of classical conditioning in which the UR is just the CR from a previous classical conditioning procedure, or the reinforced response from a previous operant conditioning procedure.[60]

ii. Additional problems for associationist theories

Mandelbaum reviews a number of problems for associationist theories of thought, two of which are particularly relevant to the primate ToM debate: learning curves and coextensionality (i.e. specifying what amounts to the "same situation"). Associative learning theories imply that associations will be acquired slowly and gradually. But behavioral data at the individual level does not show this. Gallistel et al. (2004) and Gallistel and King (2009) have re-analyzed data from animal behavior and have shown that although at the group-level, learning curves do display the properties associative learning theories would predict, no individual's learning curve

[60] Classical conditioning links an instinctual unconditioned response (UR) to a conditioned stimulus (CS). Operant conditioning reinforces a non-instinctual behavior.

has those properties. Rather, Gallistel has proposed that "learning for individuals is generally step-like, rapid, and abrupt" (Mandelbaum, SEP, 'Assoc.,' 2015, sec. 9.1, par. 1).

'Fast-mapping'[61] refers to the phenomenon of one-shot learning of a word (see footnote 16 in Chapter 3 on Rico and Chaser). Two problems emerge for the associationist in regards to fast-mapping. The first is that learning of the new word does not occur slowly, as would be predicted by proponents of gradual learning. This indicates that, "in order for word learning to proceed, the mind must have been aided by additional principles not given by the environment" (Mandelbaum, SEP, 'Assoc.,' 2015, 9.3.1, par. 1). Associationists need to explain how reasoning by exclusion can be accounted for associatively.

The second problem for associationists in regards to the phenomenon of fast-mapping is stating what the "same situation" amounts to, i.e. determining why some property is singled out as the CS when numerous stimuli are contemporaneous with the US. This is sometimes referred to as the "Credit Assignment Problem" (see Gallistel and King, 2009). According to Mandelbaum, "Associationists need a criterion to which of the coextensive properties will in fact be learned, and which not" (Mandelbaum, SEP, 2015, 9.5, par. 2). And they are yet to provide such a criterion. The "Credit Assignment Problem" demonstrates that structure is added to what is

[61] The phenomenon of fast-mapping was discovered by Carey (1978a, b; Carey and Bartlett 1978). In one of her studies, Carey showed children two otherwise identical objects that only differed in color and asked them, "Can you get me the chromium tray, not the red one, the chromium one" (Carey, 2010, 2). All of the children retrieved the appropriate tray even though none of them had any exposure to the color name 'chromium' previously ('chromium' refers to olive green). Markson and Bloom (1997) have shown that the phenomenon occurs not only for novel words, but also for novel facts.

learned, and that is problematic for the associationists' claim that learning is just the mapping of external contingencies.

Mandelbaum's explication of the variety associative theses as well as his explanation of the methods for determining whether or not some behavior is underwritten by associative mechanisms is useful in evaluating the claims made by dissenters in the animal minds debate who propose that all of the apparently intelligent behavior of nonhuman animals can be explained by associative learning and lower-order cognitive mechanisms (without reference to mental representation). According to Mandelbaum, 'association' can refer to a variety of processes: associative learning, associative structure/mechanism, and associative transitions in thought. But all associationists adhere to the belief that an organism's experience is the main sculptor of their cognitive architecture.

Associative accounts of learning reference only an organism's reinforcement history and exclude the role of inferential reasoning as guiding an animal's acquisition of new behaviors. Associative accounts of cognitive mechanisms propose that there is no structure above and beyond the associative relations between stimuli and responses. Associative accounts of transitions in thought propose that the relations amongst thoughts are acquired directly from experience, not through reasoning or inference. Mandelbaum has shown that the inference from one sense of association to another is not a given; it requires further argument and evidence.

If a cognitive architecture is purely associative, then the mechanisms it engages have no structure above the associative relation between stimuli and

responses (or between thoughts/mental representations). The fact that some skill was acquired via conditioning (i.e. associative learning) does not entail that the mind of the organism makes no contribution to the relations it stores (i.e. associative structure/mechanism) and that are instantiated in an instance of thought/thinking (i.e. associative transitions in thought). We can determine if a cognitive structure is associative (despite how it was acquired) by seeing what types of information modify, or fail to modify, behavior under control of that mechanism. If a behavior is modified by some other method than counterconditioning or extinction, then it is not underwritten by an associatively structured cognitive mechanism.

Associationism and behaviorism are distinct. And adherence to the Canon as Morgan intended it does not entail strict behaviorism. Claiming that some skill was acquired via associative learning or that it is underwritten by purely associative mechanisms does not indicate a total lack of mental representation, or even that reference to mental representations is not necessary for a successful explanation of the performance. But contemporary ABC research ignores these important distinctions and this results in incompatible accounts of the experimental data it provides.

With the relevant distinctions drawn between the varieties of associationist theses, we can now reexamine the primate ToM experiments. In doing so, I will show that it is false that all of the results of the primate studies can be explained associatively. As we will see, some of the experiments did not offer an opportunity for associative learning to have taken place, so the observed behaviors could not be a

result of conditioning. Other experiments, which did provide an opportunity for associative learning, resulted in adaptive responses on the part of the subjects that cannot be accounted for by associative mechanisms. Most importantly, though, is exposing the unwarranted jump made by the researchers of supposing that some skill could have been acquired via associative learning to the claim that that skill is carried out by an associative mechanism.

V. Wolfgang Köhler's critique of associationism

In his introduction to *The Mentality of Apes* (1925)[62] Köhler directly addresses associationism in regards to the apparently intelligent behavior of humans and other animals. He states,

> There is probably no association psychologist who does not, in his own unprejudiced observations, distinguish, and to a certain extent, contrast, unintelligent and intelligent behavior. For what is association psychology but the theory that one can trace back to the phenomena of a generally-known simple association type even those occurrences which, to unbiased observation, do not at first seem corresponding to that type, most of all the so-called intelligent performances? In short, it is just these differences that are the starting-point of a strict association psychology; it is they which need to be accounted for; they are well known to the association psychologist. (Köhler, 1925, 2-3)

Quoting Thorndike on the results of his experiments with cats and dogs, Köhler writes, "I failed to find any act that even *seemed* due to reasoning" (Köhler, 1925, 3). So Thorndike, a staunch associationist and one of the forefathers of 20th century

[62] Köhler's text reviews his research examining chimpanzees' ability to learn in the absence of explicit training, that is, whether they possess insight and are able to arrive at the solution to a problem spontaneously. Specifically, he explored the extent to which his chimpanzee subjects could construct and use tools in order to obtain out of reach food. He found that they would reliably combine two sticks into one longer instrument and that they would pile crates on top of one another in order to reach food and that they would combine these techniques when necessary.

behaviorism, recognizes the contrast between intelligent and unintelligent behavior, even though he later rejects this distinction in theory.

Köhler proposes that when we use the term 'intelligent'[63] to refer to an organism's behavior, we refer to situations in which circumstances have blocked an obvious course of action, the human or animal takes a roundabout path to obtain its goals/arrive at a solution. This is in contrast to a solution whose parts are each arrived at by chance. Köhler states,

> [O]nly that behavior of animals definitely appears to us intelligent which takes account from the beginning of the lay of the land, and proceeds to deal with it in a single, continuous, and definite course. Hence follows this criterion of insight: *the appearance of a complete solution with reference to the whole lay-out of the field.*" (Köhler, 1925, 190)

Fitting with his position as the founder of the Gestalt school of psychology, Köhler emphasizes the importance of the animal's awareness of the whole of the situation in providing them an opportunity for insight learning. Köhler's criticisms of Thorndike's experiments as well as the conclusion he draws from them are cutting. Köhler forcefully points out how Thorndike's experiments on cats and dogs in which they failed to circumvent a barrier to retrieve food on the other side failed to allow the subjects to survey the entirety of the problem situation and were thereby not conducive to insight occurring on the part of the animal subject. As with his puzzle box experiments, the animals did not have visual access to the solution so, rather, the

[63] Intelligence is a sticky topic (i.e. it attaches numerous different phenomena) and it is an ultimately ambiguous term. I prefer (and believe that it is a more productive in examining and settling disputes about the uniqueness of human cognition) to look at specific mental processes underlying apparently 'intelligent' behavior.

experimental paradigm forced the subjects to rely on chance-based trial-and-error attempts to arrive at a solution to the problem, i.e. to obtain the food reward. Köhler's experiments, in contrast, were designed and set up with an aim to allow such insight into the entire sequence of actions required to solve the problem before such actions commenced. Whether through visual access and/or prior experience in the testing environment, Köhler's subjects were not restricted to trial-and-error of chance behaviors to reach a solution.

Köhler argued that if one accepts the distinction between intelligent and unintelligent behavior in humans, then it should be obvious that his chimpanzees' behavior also lies on the side of intelligent. But following Köhler's work, the reign of behaviorism in human psychology dismissed that distinction in theory, even if not in practice. We must inquire into whether our current associative theories are able to account for the apparently insightful chimpanzees, or if their responses to the experimental manipulations put to them indicate that advanced cognitive processes are at work.

Köhler's findings cannot be explained as the result of classical conditioning because there is no reflexive behavior to have served as the UR to the US to which the CS and CR could become associated with. Nevertheless, for Köhler's chimpanzees, an associationist explanation that references operant conditioning may be able to gain traction. Köhler's chimpanzee subjects were familiar with the pen the experiments were conducted in as well as the objects that were present, so an opportunity for latent learning prior to the experimental manipulations is possible. If

the chimpanzee subjects had engaged in a variety of spontaneous behaviors in the past (i.e. chance-based trial-and-error) and received positive reinforcement by obtaining food (or whatever item they were seeking), operant conditioning might provide an explanation for Köhler's findings. But that possibility does not rule out the apes mentally representing the problem they faced before attempting a solution.

Failures could act as positive punishment if they resulted in an aversive stimulus, an injury for instance (e.g. falling from a poorly stacked pile of crates or from a pole not adequately tethered). Negative punishment (penalty) cannot explain the cessation of unsuccessful methods because the rewarding stimulus was not removed. But both forms of punishment can only account for a decrease in the given behavior—that is, in the animal learning not to perform that behavior in the future. On an associative account of learning, the successful solution must have been arrived at by chance, and the reinforcement (in this case, obtaining the bananas could act as positive reinforcement) would result in an increase of the behavior that allowed the chimpanzee to obtain it.

Kohler's experiments, despite his chimpanzees' successes, can be accounted for by either an associative account or by positing inferential reasoning on the part of his subjects. Because of this, his experimental work highlights some of the shortcomings of an ecologically valid experimental design. The subjects' familiarity with the testing environment and the objects available for them to utilize in solving the problem tasks (i.e. retrieving the food) makes it so that we are unable to rule out a prior reinforcement history for the problem-solving skills the chimpanzees engaged

in, but the diversity of behaviors he observed in his subjects does indicate a level of adaptability that may warrant a more cognitively complex explanation. Ambiguous results like this provide insight as to how to improve the experimental paradigm in a way that the findings will be more definitive in answering the question of what type of cognitive mechanism is responsible for the subject's performance.

VI. Can the primate ToM findings be explained by strict associationism?

In regards to associative accounts of apparently intelligent behavior in humans and other animals, Köhler states,

> [T]he first and essential condition of a satisfactory associative explanation of intelligent behavior would be the following achievement of the theory of association, to wit: what the grasp of a *material*, *inner* relation of two things to each other means (more universally: the grasp of the structure of the situation) must *strictly* be derived from the principle of association; "relation" here meaning an *interconnexion based on the properties of these things themselves*, not a "frequent following each other" or "occurring together" (Köhler, 1925, 219).

This applies to the primate ToM debate in that mental states cannot be observed, but can only be inferred from observable behavior.[64] Such an inference requires more than an associative relation between stimuli and behavior in order for it to be applied in novel contexts.[65] It follows that ToM cannot be explained by a pure associationist

[64] Descartes' *Second Meditation* marks the philosophical roots of ToM discussions. We can only intuit the existence of our own mind through introspection, and since we lack introspective access to another's mind we lack direct access to it. It is a 'theory' of mind because its contents are not directly observable and must be inferred from things that are observable (i.e. behavior).

[65] For example, when instructing preschool aged human children in the value of sharing we often provide negative punishment (i.e. penalty) for selfish behaviors. That is, we may remove the contested item so that neither child has access to it. But our goal is not simply to condition a sharing response in

account of cognition because it allows one to move beyond conditioned associations acquired directly from experience.

According to Mandelbaum, associative structures can only be reversed by associative learning/unlearning; "associative learning does not necessarily eventuate in associative structures, but associative structures can only be modified by associative learning" (Mandelbaum, SEP, 'Assoc.,' 2015, 4.4, par. 4). Extinction and counterconditioning are the only methods of modulating an associative structure. If a trained relation is eliminated by some other means, than it is not an associative relation, even if it was initially formed via associative learning. With Mandelbaum's method in hand, we can now examine whether the primate ToM experiments can be accounted for purely associatively, as the dissenters in the animal minds debate would have us believe.

Premack and Woodruff (1978)

Premack and Woodruff introduced the question of whether chimpanzees possess a theory of mind by performing a series of experiments aimed at determining whether or not their chimpanzee subject, Sarah, was able to "comprehend" a problem faced by a human experimenter. Sarah was tested on each of the four videotapes on

---

the child, but rather to elicit perspective taking so that the child, with time, learns to recognize and appreciate others' needs and desires so as to enable them to extend that understanding in novel contexts. The associative learning that gives rise to sharing their favored objects may then also come to facilitate additional other-oriented behaviors, such as responding to another's tears with condolences, as well as to control self-interested behaviors, such as hitting and biting. Associative learning can give rise to nonassociative trains of thought, which may or may not be underwritten by associative structures (depending on whether one accepts a simulationist or a theory-theory account of ToM).

six separate occasions, each time with a different pair of alternative solutions (i.e. the correct solution was paired with a different incorrect alternative). After selecting the photograph that depicted the solution to the problem faced by the human actor in the videotape that she had just viewed, placing it in its designated location by the television, and ringing the bell to summon the trainer, Sarah would be told either "Good Sarah, that's right" for a correct choice or "No, Sarah, that's wrong" for an incorrect choice. And regardless of her success or failure, she received yogurt, fruit, or some other favorite food at the end of each session (Premack and Woodruff, 1978, 516). It is important to note that Sarah had no experience with the testing method used, although she did have prior experience watching television, which Premack and Woodruff admit contributed to her ability to comprehend televised representations (Premack and Woodruff, 1978, 516).

There were a total of 24 trials, 6 of each of the four problem-solution sets. She was correct on all 6 trials of three of the four problem-solution sets and on the remaining problem-solution set she had 3 errors followed by 3 correct responses, which required the problem-solver to remove heavy blocks from a box before it could be moved to the location from which it could be used to obtain out of reach food. Because she received a food reward even after her incorrect choices, it is difficult to make a case for associative learning on that problem-solution set due to the lack of differential reinforcement other than the words spoken by the trainer.

Premack and Woodruff admit that they were unable to rule out an associationist explanation of the findings because it is possible that the chimpanzee

subject had seen a human engage in such behavior at some point in her past.[66] But conditioning requires reinforcement, so given the absence of differential reinforcement in the experimental paradigm, how could Sarah's ability to select the appropriate solution be acquired through associative learning as psychologists understand it? Further, Sarah's successful performance on all six trials of three of the four problem-solution sets put to her indicates one-shot learning of the experimental task, i.e. selecting the photograph that completes the sequence of actions presented in the videotape. She did not need to learn what the task being put to her was.

A stimulus-response association cannot account for Sarah's 100% successful performance on the three of the four problem-solution sets because even if the video of the actor facing a problem acted as the stimulus, the conditioned response would not be to select an image of the solution but rather for Sarah to engage in the problem solving behavior herself (for example, to use a key to open a lock). If we posit an associative transition of thoughts as underwriting her successful performance on the first trials of those problem-solution sets, then we would need to grant inner, mentalistic representations to the subject. And that is exactly what is at stake in the nonhuman animal minds debate. Granted, pure associationism need not rule out mental representation, but due to a prevalent and unfortunate misreading of Morgan's Canon, in discussions of nonhuman, nonlinguistic cognition it often goes hand in

[66] Köhler's comments on imitation as an explanation for his chimpanzee subjects' success are highly relevant to Premack and Woodruff's work with their chimpanzee subject Sarah on problem comprehension. He states, "If 'mere imitation' means imitation *without a trace of insight* into things that have been seen" (Köhler, 1925, 220), then such has not been observed in terms of complex sequences of behavior. In the human case, imitation of complex sequences involves an understating of what the action of the one being imitated means, that is, it is accompanied by insight either into the solution itself or appreciation that the sequence of actions is a solution to the problem at hand.

hand with just such rejections. Attention needs to be drawn to instances when associationism and behaviorism are being conflated in analysis of experimental results because such conflation impedes our understanding of the cognitive processes underlying the most interesting of nonhuman animal behaviors.

Povinelli and Eddy (1996)

Povinelli and Eddy (1996) performed a series of experiments in which their chimpanzee subjects never learned to selectively beg from an experimenter that could see them versus an experimenter who could not see them. Povinelli and Eddy concluded that chimpanzees do not use knowledge of what another can see to infer what that individual knows. In other words, they claim that their findings demonstrate that chimpanzees lack a theory of mind. All of the chimpanzee experiments utilized a common strategy in which the subject was presented with two human beings, one that could see them and one that could not, and was then allowed to use a behavioral begging gesture to request food from one of the experimenters.[67]

As described by Povinelli and Eddy, the mentalistic framework predicts that the chimpanzee subjects will selectively gesture toward the human being that could see them, and thereby had perceptual access to the begging gesture they were making. In contrast, the behavioral framework predicts that the chimpanzee subjects will

[67] In the blindfold treatment one experimenter had a blindfold covering their eyes and other had it covering their mouth; in the bucket treatment one experimenter held a bucket on their shoulder and other held it over their head; in the back-versus-front treatment one experimenter faced the partition separating the experimenters from the subject and the other had their back towards the partition; and in the hands-over-eyes treatment one experimenter had their palms completely obscuring their eyes while the other had their palms cover only their ears instead.

gesture to both the human that could see them and the one that could not with equal frequency. As they state,

> The behaviorist framework allows the subjects to process and use information about eyes and eye gaze and even to *learn* rules about whom to gesture in such circumstances. However, unlike the mentalistic framework, it clearly predicts that, in the situation described above, chimpanzees should initially perform at random. (Povinelli and Eddy, 1996, 26).

Preliminary training on the experimental paradigm consisted of the chimpanzee subjects being taught to put their hand through the one of two holes in a transparent screen that an experimenter was standing in front of in order to receive a food reward. During this portion of the experimental procedure, the experimenter either "visually focused on a neutral spot adjacent to the testing unit in order to minimize any potential cues that he might give the animals" (Povinelli and Eddy, 1996, p. 29) or "fixed their gaze on a small target that was positioned exactly midway between the two response holes" (Povinelli and Eddy, 1996, p. 30). They state,

> For all testing trials, including all standard, baseline, and treatment probe trials, the experimenters fixed their gaze on the target midway between the two response holes (33) […] [W]e reasoned that, if chimpanzees truly understand that seeing connects someone to the external world, it should not matter whether the subjects see the experimenter looking at them directly or where they are about to respond. (Povinelli and Eddy, 1996, p. 35)

It took the chimpanzee subjects between 250-750 training trials to achieve the 38 of 40 correct response criteria at gesturing at the hole in front of the experimenter. Because the experimenters were not actually making eye contact with the subjects during this extensive training on the experimental task (i.e. begging from an experimenter that could see them), it is possible that this training actually had the

result of untraining a relation between eye-gaze and attention/visual awareness in the chimpanzee subjects. If the chimpanzee subjects were unintentionally trained to beg from someone facing them but not looking at them, then both the blindfold and the bucket conditions would no longer provide the subject with a salient cue as to whether they could expect a response from the experimenter.

In contrast to their poor/chance performance on the treatments involving nonnaturalistic objects (i.e. blindfold and bucket), five of the six subjects performed perfectly on the four trials of the treatment involving one experimenter facing them and the other facing away. Povinelli and Eddy admit that the subjects did not need to learn to selectively beg from the experimenter facing them since they did so on the first trial. Further, "the subjects showed no evidence of improving across the four trials of the other treatments that were administered" (Povinelli and Eddy, 1996, p. 43). The subjects were given a food reward for gesturing to the experimenter who could, in principle, see them and were not rewarded if they gestured toward the experimenter that could, in principle, not see them.

If associative learning is the explanation proffered to explain all non-reflexive animal behavior, then why did the chimpanzee subjects never acquire an association between having access to an experimenter's eyes and receiving a food reward? How can associationism account for this failure to learn on the part of the chimpanzee subjects? The experimental findings neither support the associative account nor refute the cognitive account of the processes underlying the chimpanzees' performance. Granted, associative learning procedures are not always successful, but

the failure of what was in fact an operant training paradigm cannot be taken as evidence of an associative structure/mechanism. Additionally, if the training trials actually had the result of decoupling an associative relation between the experimenter seeing them and responding to their begging gesture, by providing a food reward over hundreds of trials in which the experimenter was looking at a neutral spot and not at the begging subject, what the test trial results indicate is that the association between eye-gaze and awareness may have been unintentionally counter-conditioned.

Povinelli and Eddy clearly contrast results that would be the outcome of the operation of a ToM from those that can be explained by associative learning. They state,

> [W]e designed our studies in order to allow for a very sensitive diagnosis of whether the animals possessed immediate dispositions to act in a fashion predicted by a theory of mind view of their psychology or whether their successful performances could be better explained by learning theory. (Povinelli and Eddy, 1996, v)

In doing so they reveal the unwarranted transition they make between associative learning to associative structure. As explained by Mandelbaum, the basis of some skill in associative learning does not indicate that that skill is being carried out by associative mechanisms. Simply demonstrating that some animal's experiential history provided an opportunity for latent learning of a task is not sufficient for claiming that the ability acquired is underwritten by purely associative cognitive mechanisms.

Hare, Call, Agnetta, and Tomasello (2000)

Hare and colleagues' "breakthrough experiments" involved pitting a subordinate chimpanzee and a dominant chimpanzee against one another in a competition over food in a situation where only the subordinate had knowledge of the location of a second piece of food. The results showed that the subordinate reliably avoided the food the dominant could see and pursued the food the dominant could not see in the experimental conditions in which the subordinate had better visual access to the food than the dominant.

Ten adult and sub-adult chimpanzee subjects were first tested for dominance relations using a food competition test. A single pair of animals was put into a cage together and experimenters placed a piece of fruit equidistant between them. The subjects who obtained the food in the presence of the conspecific were assessed as dominant in that pair. Seven of the ten chimpanzees were deemed subordinate to someone else in the group and the other three were used only to obtain data on those seven subordinate subjects. Therefore, some subjects served the role of both subordinate and dominant depending on which conspecific they were paired with in a given trial.

In the pilot experiment, subjects were tested in pairs consisting of one dominant and one subordinate chimpanzee. There were four testing conditions, defined by the location of the subjects and the location of the target food. In all cases the subordinate had visual access to the baiting procedure. In the first condition, the subordinate and dominant could see one another through an open doorway and an

experimenter introduced food into the subordinate's cage in a way that the dominant could neither see the baiting nor the food once baited. In the second condition, both the subordinate and dominant were in the dominant's cage and the food was introduced equidistant between them, so both subjects had visual access to the baiting and the food throughout. In the third condition, the subordinate was standing inside the doorway connecting the two cages with the dominant in its own cage and the food was placed in the subordinate's cage next to the wall separating the two cages, so the dominant could not see the baiting nor the food once baited. And in the fourth condition, the subordinate was inside the doorway while the dominant was in its own cage and the food was placed in the dominant's cage next to the wall. Across the four conditions, the subordinate obtained a significantly greater proportion of food in the conditions in which she had exclusive visual access to the food, i.e. in the first and third conditions, in which the dominant could not see the baiting procedure nor the food once baited.

In the first experiment of the series, referred to as "The Wall Test," two target foods were available with at least one always in plain sight of both subjects and equidistant to them. The variation across three conditions was where the second piece of food was placed. In the first condition, one piece of food was in the doorway visually accessible to both subjects and the other piece of food was in the dominant's cage next to the wall so that it was visually inaccessible to the subordinate. In the second condition both pieces of food were placed in the door ledge separating the two rooms, so both subjects could see them. In the third condition, one piece of food was

placed in the doorway and the other was placed in the subordinate's cage next to the wall, so the dominant could not see it. Results showed that the subordinate obtained a significantly larger proportion of food in the Subordinate-Door condition than in either the Dominant-Door or Door-Door condition. Further, of all the food obtained by the subordinate in the Subordinate-Door condition, 83% came from their own cage (Hare et al., 2000, 775). Equally interesting, the researchers found that in the Dominant-Door condition, dominant subjects first took the food that they could both see and then after that took the food in their own cage that only they had visual access to.

In order to explain the results of Experiment 1 via associative learning we must posit a learning history for the subordinate subject in which they had initially, by chance, retrieved food from their own side of a barrier on which a dominant was on the opposite side (positive reinforcement) and/or had unsuccessfully attempted to retrieve food that was equidistant between themselves and a dominant (negative reinforcement). Further, an associative learning explanation of the dominants' behavior in the Dominant-Door condition (i.e. first retrieving the food visually accessible to both the dominant and the subordinate and then retrieving the food that only the dominant could see) requires us to postulate a different learning history that would lead them to first take the food visually accessible to both themselves and the subordinate before retrieving the food that only they could see. And since some of the testing subjects played both roles, we need to posit additional associative

mechanisms that switch on the alternative responses depending on the dominance relations between themselves and the conspecific they are interacting with.

Experiment 2, "The Tyre Test," was designed to rule out alternative explanations of the previous experiments by introducing two key modifications. First, the physical barrier blocking visual access (i.e. the wall) was replaced with a tire that was equally physically accessible to both subjects, although only the subordinate knew the location of the hidden piece of food. And second, the two pieces of food were placed in a single cage that both subjects had equal physical access to, so the dominant was present when the subordinate attempted to obtain the hidden food. The three testing conditions varied by the location of the food, and thereby by whether the dominant had visual access to it once the door on their cage was raised. In each trial the subordinate had visual access to the baiting process while the dominant did not. In the first condition one piece of food was placed on top of the tire, so visually accessible to both the dominant and the subordinate, and the second piece of food was hidden inside the tire, so visually inaccessible to both subjects (although the subordinate had witnessed the baiting procedure). In the second condition, one piece of food was placed at a distance alongside the tire, so both subjects could see it, and the other was placed directly behind the tire from the dominant's point of view, so only the subordinate could see it. And in the third condition both pieces of food were placed on top of the tire, so visible to both subjects.

Throughout the baiting procedure the subordinate's door was raised enough for her to see the experimenter placing the food, but not enough to emerge from the room. After the experimenter exited the testing room, the dominant's door was raised so that she could also see the physical situation in the room and the subjects could see each other. After both subjects had looked through their doors, both doors were raised simultaneously allowing the subjects to enter into the middle room where the tire and food were located. Results showed that subordinates reliably obtained food only in the conditions in which they had exclusive visual access to its location. Further, "some subordinates also behaved in strategic ways to avoid detection when reaching for hidden food" (Hare et al., 2000, 778). Four subjects on seven different occasions approached the tire but then waited to retrieve the hidden food until the dominant had moved away. Three subjects, on one occasion each, used even more proactive strategies. One subject waited until the dominant had turned its back to retrieve the food, even though the dominant was still close by. And two subjects gave active communicative signals to the dominant (a greeting and a sexual presentation) in an apparent effort to keep the dominant on its own side of the tire, and then used their bodies to block the dominant's line of vision while retrieving the hidden food.

Although the learning history required for results of Experiment 1 could explain the subordinates' behavior in the condition in which the second piece of food was placed on the subordinates' side of the tire as well as when both pieces of food were placed on top of the tire, an additional learning history is required to explain the subordinates' behavior in the condition in which the second piece of food was hidden

inside the tire. In this condition the food was visually inaccessible to both the subordinate and dominant once hidden, but the subordinate witnessed the baiting and the dominant did not. So the subordinate would need to have by chance trial-and-error been positively reinforced in the past on enough occasions to condition the behavior of retrieving food that only they had seen someone else hide. And for the three subjects that engaged in pro-social behaviors to deceive the dominant must also be posited to have additional learning histories to explain that behavior.

Experiment 3, "The Occluder Test," was designed to rule out the possibility that in Experiment 2 the subordinate subjects were simply reacting to the dominant's intention movements. To guard against this, in this experiment the subordinates were given a very small temporal head start. The researchers state," The question is whether they would go immediately for the food to which they had exclusive visual access, or whether, alternatively, they would go for the food openly visible to both contestants" (Hare et al., 2000, 779). As with Experiment 2, the testing situation involved three rooms connected by doorways, with the subjects placed in the two outermost rooms at the outset and the food placed in the middle room. Neither subject could see the baiting procedure because their doors were closed. Once the experimenter left the middle room, both doors were raised only enough for the subjects to see inside the middle room and each other for a few moments before they were opened and the subjects were allowed to move through them to the center room.

In the first condition, two pieces of food were placed in the middle room equidistant from the subjects' doorways and visible to both subjects. In the second

condition a piece of PVC pipe acted as an occluder, blocking the dominant's visual access to one of the pieces of food with another piece of food visually accessible to both subjects. In the third condition, both pieces of food were blocked from the dominant's vision by occluders. Subordinates obtained the most food in the Hidden-Hidden condition, followed by the Hidden-Visible condition, and the least in the Visible-Visible condition. The researchers also note that in the Hidden-Visible condition, despite being given a head start, the subordinates started out for the hidden food on 73.4% of trials (Hare et al., 2000, 780). Further, on four separate occasions involving three individuals the subordinate waited to take the hidden food until the dominant had moved away. And one subordinate used her head start to race to the dominant's doorway and greet her, effectively keeping the dominant in her own cage. The subordinate then managed to get both pieces of hidden food.

An associative explanation of the findings requires us to postulate learning histories that involve successful chance-based attempts of obtaining food behind (from the dominants' point of view) much smaller objects than a barrier, only big enough to block the object hidden. This is unlike the situation in Experiment 1 where, from the position of the food behind the wall, the conspecific is not visible to the subject. That is, in the third condition of Experiment 1, once the subordinate arrived at the location of the second piece of food, the subordinate lacked visual access to the dominant and the dominant lacked visual access to the subordinate. But in Experiment 3, the subordinate and the dominant maintained visual access to one another throughout; it was only the dominant's visual access to the food that was

blocked. So the learning history from Experiment 1 cannot also serve to explain these results, although the learning history for the second condition in Experiment 2 may be able to do so because in both Experiment 2 and Experiment 3, the dominant was present and visually accessible when the subject attempted to retrieve the food. Also, as with Experiment 2, the deceptive behaviors of three of the subjects also require additional learning histories to be accounted for associatively.

Experiment 4 was performed as a guard against a possible alternative "intimidation hypothesis" explanation of the results of Experiment 3. As the researchers explain, "It is thus possible that they [the subordinate] established the dominant's visual gaze direction and then, based on past experience in which they tried unsuccessfully to get food the dominant was looking at, chose the food the dominant was not looking at" (Hare et al., 2000, 781). The set-up was the same as in the previous experiment, except that the dominant's door was completely closed during the subordinate's selection period, so that subordinate could not base their choice on the dominant's eye gaze or other behavioral cues. Results replicated the findings of Experiment 3, ruling out the intimidation hypothesis.

The final experiment in the series, "The Transparent Barrier," replicated the experimental situation of Experiments 3 and 4 but utilized a transparent barrier rather than an opaque one. Additionally, two rounds of trials with different delays between the release of the competitors were performed. In the first round, the dominant's release was delayed until the subordinate had made its choice (as in Experiment 4). In the second round, the dominant's release was delayed only until the subordinate

had chosen a direction (as in Experiment 3). In the first round, with the longer delay, the subordinates showed no preference for the food behind the transparent barrier, but with the shorter delay the subordinates showed a small preference for the food out in the open. The researchers state, "We have no ready explanation for the subject's preference for the food in the open in the second round of testing, but the main point is that they did not prefer the transparent barrier and so they clearly did not think that the plastic bottle was occluding, or in any way hindering, the vision or behaviour of the dominant" (Hare et al., 2000, 783). The researchers go on to note,

> Indeed it is notable that our chimpanzees did not have to be trained in any procedures nor did they have to interact with humans at any time during the testing, which would seem to provide a priori evidence that our paradigm is a very natural one in which to assess primate social-cognitive skills. […] Chimpanzees' strategically appropriate behaviour with the transparent barrier was perhaps especially important since they had not had much experience with transparent objects previously and so could not have had many opportunities to learn specific contingencies between these objects and the behaviour of their groupmates. (Hare et al., 2000, 783)

So, for Experiment 5, it is unlikely that the subjects had had enough experience with transparent occluders for an associative learning history to have even taken place. But because of the different behavior of the subjects depending on whether the occluder is transparent or opaque, such a learning history is necessary for an associative account to explain the findings.[68]

[68] Most of these studies focus on the absence of associative learning. It seems that both the supporters and dissenters in the debate over primate ToM equate associative learning to purely associative structures/mechanisms, and vice versa. Hare and colleagues appear to be making the unjustified (according to Mandelbaum) leap from a lack of associative learning on a particular problem set/task to an absence of associative structures/mechanisms able to carry out the solution to that task. But according to Mandelbaum, we can acquire an associative structure from non-associative learning. For example, reasoning by exclusion (as occurs in fast-mapping) can give rise to an association between a

Across all of the various experiments, the researchers found that if a subordinate recognizes that a dominant fails to see the baiting of food, that the subordinate will retrieve it. This holds true even when the subordinates are given a head start; the subordinate avoids food the dominant has seen. The researchers also demonstrated that subordinates avoid food behind a transparent barrier but not an opaque one.

Even more noteworthy for our present purposes, "The fact that the same individuals adopted different strategies depending on the role they played in the experiment, subordinate or dominant, suggests that they were not following some blind behavioural contingencies or rules" (Hare et al., 2000, 783). The researchers conclude that, although they do not believe that chimpanzees understand visual experience in the same way that humans do, their findings demonstrate that neither non-cognitive physiological reflexes nor behavioral conditioning can explain the subjects' adaptability to different dominance relations demonstrated by the experimental results. They state,

> As is often the case in post hoc behavioural explanations, to account for our findings this cognitively weak hypothesis would have to posit different sets of learned contingencies for virtually every experiment (and indeed different sets of contingencies for the subordinate and dominant roles in each experiment), including the transparent barrier test, which involved an 'occluder' with which the subjects had had very little experience. Such post hoc scenarios, based on no actual observations of individuals' behaviour, seem highly unlikely. (Hare et al., 2000, 784)

---

word and a property or object. When instructed to "Get me the chromium tray," and you know that one tray is red, another blue, and the last a color you have never heard named, reasoning by exclusion guides one's choice of the correct object. But this non-associative learning of the color name 'chromium' in turn results in the acquisition of an association between the word 'chromium' and the particular color of green that the tray is.

Experiment 1 requires us to posit numerous learning histories for each individual subject: a learning history for the subordinate role, another learning history for the dominant role that causes them to perform the opposite behavior, as well as a learning history for each dominance relation they are party to. Experiment 2 requires an additional learning history for hidden objects. Experiment 3 necessitates another learning history for dealing with small occluders that block the visual access of the dominant to the food, but do not block the subordinates' visual access to the dominant. And Experiment 5 requires a learning history that dictates different behavior when an object is behind a transparent occluder than when the occluder is opaque, which is unlikely to have taken place in the subjects' past given the absence of transparent barriers in their environment.

Taken together, this series of experiments displays the plasticity of the chimpanzee subjects' behavior and demonstrates that it is adaptive enough to be the product of reasoning. Associations are not easily extinguished. So even if such associations were formed in the earlier experiments in the series, the fact that the chimpanzees exhibit different behaviors in the different experimental conditions that have the same goal and adapted to the circumstances, as well as different behaviors in the same experiment depending on whether they are in the role of subordinate or dominant, suggests a non-associative explanation.

In a second set of experiments a subordinate watched through an open doorway as an experimenter placed a piece of food on her side of a doorway while a dominant was on the opposite side of that barrier, sometimes with visual access to the

baiting and sometimes without such access (the dominant's door was either open or closed, and the subordinate could see which was the case). Unlike the first set of experiments, there was only one piece of food available. Results indicate that the subordinate was aware of whether the dominant had witnessed the baiting moments before. That is, if the dominant's door had been closed during the baiting process, the subordinate retrieved the food and if it had been open during the baiting process, the subordinate avoided the food even though it was at that time only visually accessible to the subordinate.

Povinelli together with Giambrone (2001) reject the conclusions drawn by Hare et al. and posit instead that subordinates may simply prefer food next to a barrier over food out in the open and that their selection of the food behind the barrier in the first set of experiments could be explained without making reference to the subordinate's awareness of what the dominant had perceptual access to. So Tomasello et al. (2003) ran a control condition with a subordinate in a non-competitive task and found that the subjects did not prefer food next to barriers. Further, in the second set of experiments this explanation is not sufficient to explain the results because there was only one piece of food and it was always next to a barrier. Despite this, the subordinates' choice was determined by whether or not the dominant had witnessed the baiting.

In their natural environment dominants take all the food and punish subordinates who challenge them, so the finding that subordinates prefer food that the dominant has not seen is not remarkable. The requisite associations could have been

acquired in the animal subjects' ontogenetic history, so perhaps we cannot rule out that the behavior was acquired via latent conditioning. But as Mandelbaum has demonstrated, the fact that some behavior is the result of associative learning does not entail that what was learned is necessarily associatively structured. Hare and colleagues seem to recognize this when they state,

> During their ontogenies, as they follow the gaze and attempt to predict the behaviour of others in many situations, individuals may learn many additional things about the relation of their groupmates' visual access to objects in the environment and its implications for their (their groupmates) subsequent behavior (…) During these learning experiences the observer may sometimes see that the other individual is afraid or is excited about something (and so avoiding or approaching it) which the observer cannot initially see (e.g. because someone or something is occluding its view) but which it then comes to see later. *Such situations provide experience for making the connection between visual access and the behaviour of others in various social contexts*. (Hare et al., 2000, 784, italics added)

And later,

> It is important to emphasize that our mixed explanation is not equivalent to a behavioural conditioning, noncognitive explanation. Even though it involves learning, it may be construed as a cognitive form of learning which leads to real understanding and insight, as expressed in knowledge that is flexibly displayed in behaviour. (Hare et al., 2000, 784)

According to the researchers, it follows that blind conditioning and theory of mind are not the only explanatory alternatives. On their third alternative, chimpanzees' insight into social problems is similar to their insight into physical problems like spatial reasoning and tool use; "with this insight in all cases depending to some degree on personal experience with the objects and activities involved." And further, "On a daily basis chimpanzees find themselves in novel social situation for which they

devise novel strategies, or adapt known strategies, based on a knowledge of the structure of the social problem" (Hare et al., 2000, 784). A mixed account of the cognitive mechanisms underlying the chimpanzee subjects' behaviors across the various experimental manipulations is more parsimonious, in numerous ways, than is a purely associative account of the findings. Supposing that direct experience in which a subordinate is rewarded for attempting to obtain food that a dominant did not have visual access to gives rise to a non-associative mechanism that allows them to extend an understanding of the relation between visual access and awareness to novel situations better accounts for the findings than does an ad hoc, cognitively simpler explanation that requires us to posit numerous unverifiable learning histories for all of the conditions of each of the experiments, as well as across the various dominance relations for each of the subject pairs.[69] Whereas the latter explanation may purport to be psychologically/qualitatively parsimonious (in terms of the complexity of the cognitive processes being proposed to account for the findings), the former explanation is both quantitatively parsimonious (in terms of the number of distinct cognitive mechanisms being postulated) as well as evolutionarily parsimonious (in regards to the phylogenetic proximity of chimpanzees to human beings).

Call, Hare, Carpenter, and Tomasello (2004)

[69] See footnote 48 for Colin Allen's comments on transitive inference and dominance relations. Some dominance relations are acquired via direct experience, but an inferential cognitive mechanism that allows an animal to deduce the consequences of those relations allows them to extend knowledge acquired directly to novel relations.

Call, Hare, Carpenter, and Tomasello (2004) performed a series of experiments which demonstrated that chimpanzees are more impatient and apparently angrier with experimenters that are unwilling to give them food than experimenters who are unable to do so. The researchers propose that the results indicate that chimpanzees can discriminate between intentional and accidental actions.

In this series of experiments, Call and colleagues demonstrated that their chimpanzee subjects spontaneously (i.e. without training) behave differently depending on whether a human experimenter is unwilling versus unable to give them food. They propose that this indicates that chimpanzees know more about intentional actions than had been previously demonstrated. In an effort to improve upon previous experimental paradigms (specifically, Povinelli and Eddy's 1996 series of experiments), the researchers used chimpanzee subjects who had no previous training in a cognitive testing environment, they included multiple conditions to minimize the chance that the chimpanzees' performance was due to a superficial clue rather than the intent of the actor, and they utilized apes' natural responses.

The testing subjects were 12 chimpanzees between the ages of 4-26. All 12 subjects had been moved to the Leipzig Zoo within the last 6 months and none had completed any cognitive experiments prior to the study. During the test trials they did not receive any food rewards. In the unable conditions, the experimenter could not transfer food to the subject either because something prevented it or because they were distracted. In the unwilling conditions, the experimenter simply refused to transfer the food.

In Experiment 1 an experimenter began giving food to a chimpanzee subject through a hole in a Plexiglas window, but the food transfer was delayed, either because the experimenter refused to give the subject the food or because they were unsuccessful in doing so. In the testing sets, two unable conditions and one unwilling condition were presented. The experimental question was whether the chimpanzees would behave differently across the various unwilling conditions from their behavior across the various unable conditions.

Three trios of conditions were presented in each test trial. The "tease trio" included an unwilling tease, unable clumsy, and unable blocked hole condition. The "refuse trio" included an unwilling refuse, unable distracted, and unable can't see (the experimenter unwittingly dropped a grape in a location they could not see it from so was, from the chimpanzee's perspective, not aware of the grape's location) conditions. In the "eat trio" there was an unwilling experimenter eats the grape condition, unable search condition (the experimenter set a grape on the table but then continued searching in the bucket, as if they had forgotten that they had already retrieved the grape from the bucket), and an unable stuck condition (the experimenter unsuccessfully attempted to get a grape out of a transparent tube that it was stuck in). The refuse trio did not show significant differences in the subjects' behavior across conditions, but both the tease trio and the eat trio did. The researchers found that the chimpanzee subjects produced significantly more behaviors in the unwilling conditions than in the unable conditions when the experimenter physically acted on

the grape (the tease and eat trios). The subjects also left the testing station earlier in the unwilling conditions than in the unable conditions in those trios.

Experiment 2 was designed to determine whether the findings from Experiment 1 could be due to something other than the chimpanzees' attempt to communicate with the experimenter. The researchers utilized a non-social test in which they compared a condition in which the experimenter left the room after placing a grape on the table in front of the Plexiglas window with the unwilling refuse and the unable can't see conditions from the previous experiment. They also presented the subject with the unwilling tease, unable clumsy, and unable stuck conditions in test trials both with and without food. The subjects' behavioral rate was significantly lower in the no experimenter condition and the subject left the testing station earlier than in the other two conditions. They found no significant difference in the subjects' behavioral rate across conditions in the food versus no food trials, but they did find significant differences in the subjects' latency to leave the testing station. The controls introduced in Experiment 2 indicate that the chimpanzees were attempting to communicate to the experimenter and not simply displaying their frustration.

The chimpanzee subjects in these experiments received no training prior to the test trials, so a purely associative explanation of the results would require us to postulate prior latent learning histories for each of the five experimental manipulations. In regards to an associationist explanation of the findings, the researchers state,

> [I]f chimpanzees were using their previous experience of E's actions to decide how to react, they would have had to have a separate learning history for each of the five conditions in which they discriminated successfully. This is unlikely because some conditions, at least, arguably were novel to the chimpanzees and because these chimpanzees had little experience with experimenters or testing in general because they were new to the facility. Note that chimpanzees could not have developed such an expectation during the test because they were not differentially rewarded in the experimental conditions and there was no training involved. (Call et al., 2004, 496).

So, although Hare et al. (2000) and Call et al.'s (2004) findings can, in principle, be accounted for purely by associative learning, such an explanation would surely be dismissed in the human case. Even if we grant associative learning as able to account for the findings, by postulating non-associative mechanisms as arising from that learning we arrive at a far more reasonable and parsimonious account of the totality of experimental results. As Call and colleagues state, "the current study suggests that chimpanzees do not simply perceive the behavior of others, they also interpret it" (Call et al., 2004, 497). And that ability to interpret such behavior may ultimately have been based in associative learning processes, but the thing learned is not associatively structured because of the ways it extends to novel situations the animals find themselves in.

Flombaum and Santos (2005)

Flombaum and Santos (2005) extended Hare et al.'s findings to non-ape primates. In a series of six experiments the researchers found that their free-ranging rhesus macaque subjects reliably retrieved a grape from an experimenter who could not see them and avoided the experimenter who could see them. The researchers

state, “Success in this situation depends on more than mere gaze following; subjects must spontaneously use information about direction of an individual’s gaze to make a task-relevant decision” (Flombaum and Santos, 2005, 447). There were no learning trials necessary for this behavior to initiate; a single experimental trial was performed on an individual subject so the abilities demonstrated were available to the subjects in the absence of direct training.

All six of the experiments used two human experimenters, dressed identically and matched for physical attributes. The experimenters simultaneously presented the monkey subject with a grape secured to a white foamcore platform that they placed on the ground. In Experiment 1 subjects were given the choice of approaching an experimenter that was facing them or an experimenter that was facing away, and the subjects reliably approached the experimenter facing away. In Experiment 2 the experimenters placed their platforms on the ground to their side and then either faced the platform or faced away from it. Subjects reliably chose to approach the experimenter facing away from his platform. In Experiment 3, after placing their platforms on the ground in front of them (facing the subject), one of the experimenters turned their head 90° away while the other kept their head facing forward. In Experiment 4, after placing their platforms on the ground in front of them, one of the experimenters averted their eye gaze 45° to the side while the other continued gazing forward. In Experiment 5, after placing their platforms on the ground in front of them, the experimenters held up large opaque barriers (20 x 80 cm). One of the experimenters held it so as to cover their entire face and the other

held it in front of their chest. Experiment 6 was like Experiment 5 except that the opaque barriers were much smaller (6 x 20 cm) and one experimenter held it over his eyes while the other held it over his mouth. As the researchers state, "Even without training, our subjects knew to attend to the specific features of the competitor's posture that determined what they could and could not see: the direction of their eyes" (Flombaum and Santos, 2005, 448). Across all six experiments, the monkey subjects reliably approached the researcher that could not see them.

Flombaum and Santos consider a deflationary account of the results of their experiments according to which the subjects' success did not depend on reasoning about the mental states of the human experimenters, but by simply following a rule to avoid the experimenter looking forward. But they conclude that such an account is unlikely given the results of experiments 4, 5, and 6. They also consider another non-mentalistic explanation according to which the subjects simply avoid the experimenter whose body posture makes him more likely to respond to their approach. But they state,

> Note that a rule such as this could potentially explain any ToM-like behavior without the explicit representation of the mental state of another individual (…) We would like to argue that, in this context, applying such a rule successfully is precisely the point and, indeed, qualifies as reasoning about the perceptions of others. This is because such rules *should not* apply successfully in all contexts. (Flombaum and Santos, 2005, 450)

As they note, although such a rule would work in a competitive foraging situation it would not lead to success in many other contexts, for example, when attempting to attract potential mating partners or in caring for young. ToM allows an organism to

predict the behavior of another individual by inferring their internal mental states from their external observable behavior *together with* the context in which that behavior is being performed (i.e. the contents of that other individual's perceptual awareness).

Within the experimental procedure, the monkey subjects were never punished (i.e. positive punishment); no aversive stimulus followed taking the grape from the human experimenter that could see them. It is also not possible for the monkey subjects to have acquired the associations during the experiment because each subject was only used for a single trial. But as with the naturalistic experimental paradigms utilized with the chimpanzee subjects, we cannot rule out the behaviors' basis in latent associative learning in their natural environment, but in this case those previous experiences are not as far-fetched. Perhaps the monkey subjects refrained from taking the grape because an association had been formed via previous interactions with conspecifics between seeing another's eyes and not receive the food/reward (i.e. penalty/negative punishment).

Santos, Nissen, and Ferrugia (2006)

Santos, Nissen, and Ferrugia (2006) performed two additional experiments on the same population of free ranging monkeys which demonstrated that monkeys will reliably take a grape from a silent container that does not alert the human experimenter to what they are doing and will avoid the noisy container. In the first experiment, the researchers began each trial by displaying the auditory properties to

the subject. They opened the lid of the container, removed the grape and displayed it to the subject, returned the grape to the container, shaking the lid throughout (whether it had intact jingle bells or altered ones that caused it to be silent). The experimenter then retreated approximately 2 meters back, squatted down, and put their head between their knees so that they could not see either the subject or the containers. The subjects reliably approached the silent container.

An alternative explanation discussed by Santos et al. is that the subjects avoided the noisy container simply because they were afraid of it. Experiment 2 was performed to address this possible objection to the conclusions drawn from the results of Experiment 1. In the second experiment, after retreating from the containers and squatting down, the experimenter continued to look in the direction of the subject. Of the 16 subjects that approached one of the two containers, only 5 chose the silent container. These results rule out the alternative explanation of Experiment 1's findings. Taken together, these experiments indicate that, "Our subjects seemed to understand that the competitor's ability to hear their approach was irrelevant if he had already seen (and thereby already knew) that they were approaching" (Santos et al., 2006, 1179).

Povinelli and colleagues have proposed that primates' successful performance in ToM tasks can be explained by positing a sophisticated form of behavior-reading in the absence of mind-reading abilities. Santos and colleagues respond to the claim that "monkeys could succeed in mind-reading tasks by reading and abstracting competitors' behaviors without any knowledge of their mental states" by contending

that "Any behavior-reading account relies on primates having a historic link between some aspect of a competitor's observable features and (e.g. the direction that their eyes are pointing) and his future behavior" (Santos et al., 2006, 1180). But they point out that such a historical link is absent in this case;

> Because monkeys in this population have never had the possibility to test how jingling sounds affect a human competitor's future actions, they could not have built up the experiences needed to make behavioral predictions about the competitor's likely response, which would be required for a behavior-reading account of our results. (Santos et al., 2006, 1180)

VII. Conclusion

Mandelbaum distinguishes three kinds of associations: association as a type of transition between thoughts, as a mental structure, and as a learning procedure, and argued that it was invalid to simply infer from one sense to another. He states, "Distinguishing between these senses of 'association' is important because they illuminate reasonable theoretical possibilities" (Mandelbaum, 2014, 18). One can conclude that some response is the product of long-term exposure to stimulus *A* being followed by stimulus *B* and still deny that the response itself is associative in any interesting sense.

In cognitive science, the associative account has largely been dismissed with. It strangely remains in social psychology. As Mandelbaum points out,

> Not many psycholinguists take associative structure to be the only type of representational structure. This is because one really can't do psycholinguistics (never mind generative semantics or syntax) without, at a minimum, structures that take truth-values, and because

> associations aren't truth-apt, they cannot serve that role. (Mandelbaum, 2014, 18)

Mandelbaum also notes that associationism is still thriving in the animal cognition tradition (Mandelbaum, SEP, 'Assoc.,' 2015, section 8). Mandelbaum attributes the proliferation of dual-process theorizing to this oddity in which social psychological theorizing retains associationist theses while the majority of the rest of the cognitive sciences quantify over propositional structures. So it is understandable that in the animal cognition literature, folk-psychological processes are often accounted for by associative processes.

Hare and colleagues' set of experiments utilizing a competitive rather than a cooperative task is considered a "breakthrough" because it required the subject to engage in an ecologically valid task. As I've argued elsewhere, these naturalistic experimental paradigms are least likely to provide incontrovertible evidence of advanced cognitive abilities in nonhuman, nonlinguistic animals. Mandelbaum's discussion of the ways that associative learning can be dissociated from associative structure/mechanisms should make us consider the possibility that even species-typical traits may be underwritten by non-associatively structured mental states/contents, even if those contents were acquired via associative learning. Nevertheless, the experiments hailed in the ABC literature for their reliance on ecologically valid problem tasks are actually those for which it is most difficult to rule out once and for all their basis in associative learning. And because the focus in the association/cognition debate over nonhuman animal minds has largely focused on learning histories, the stagnation of the debate remains.

Associative learning can give rise to non-associative structures (for example, discussed previously in section II subsection ii, one can gain an associative structure that has a proposition as one of its associates or a propositional structure can result from associative learning). Those non-associative structures can then become associative inferences (as discussed in section II subsection iii on associative transitions). It is for this reason that naturalistic experimental paradigms are ill-suited for demonstrations of advanced cognition in nonhuman animal subjects. Knowing that some skill is the product of associative learning does not entail that it is associatively structured and knowing that some form of thought is carried out by associative transitions between thoughts does not indicate that the mechanism itself is associatively structured or that it was acquired via associative learning. When examining only species-typical behaviors of an organism we are left with the possibility that the trait arose as a result of associative learning in the individual's ontogenetic history, even if the mechanism itself is not associatively structured. Further, the additional possibility that over time the relata of that non-associative structure become associatively related also remains. But none of this indicates an absence of higher-order cognitive mechanisms playing a role in the deployment of the given skill.

In ABC research, behaviorism and associationism are all to often conflated. Associative transitions in thought are absent from associative accounts of nonhuman animal behavior due to a continued blind adherence to the primary methodological principle of ABC research, i.e. Morgan's Canon. In addition, positing associative

learning as responsible for the acquisition of some behavior is taken as direct evidence of an associative structure/mechanism underlying the behavior. Both of these things are unwarranted. So associative accounts of the primate ToM studies fail on two counts:

1) Reference to stimulus-response associations alone (vs. associative transitions in thought) resulting from conflation of associationism and behaviorism eliminates appeal to the associationist theses that may be of most value to those attempting to explain all nonhuman animal behavior by pure associationism (i.e. learning, structure, and transitions). But doing so would require one to take the position in the nonhuman animal minds debate that internal representation (i.e. thought) is not restricted to the human species.
2) Mandelbaum has shown that we cannot infer from associative learning to associative structure without further evidence and argumentation and in the case of the primate ToM literature: (a) associative structures cannot account for the adaptability and novel behaviors displayed by the chimpanzee subjects (and cannot account for Rio's ability to infer equivalence relations on novel stimuli sets), and (b) some of the primate ToM studies provided their subjects with no opportunity for associative learning on the problem task, so it is also possible that even if the mechanism underlying the chimpanzee and monkeys' successful performance is associatively structured, it must have arisen from non-associative learning (reasoning by exclusion, for example).

Some skill can be acquired via associative learning but still not be underwritten by an associatively structured cognitive mechanism. Further, a non-associatively structured mechanism can give rise to an associative transition with time and repeated application. It seems reasonable to posit that for some advanced cognitive traits, like formal reasoning procedures, that they were acquired via associative learning on a set of exemplars but that the thing acquired is a non-associatively structured mechanism (i.e. conceptual understanding of the problem).

Experimental manipulations/paradigms that ensure the presence of non-associative mechanisms are our best bet in settling the debate over the status of advanced cognition in nonhuman, nonlinguistic animals. Schusterman and colleagues' work with their California sea lion subject, Rio, represents such an experimental case. In the case of Schusterman et al.'s demonstration of Rio's acquisition of an equivalence concept, it is not possible to develop a purely associative account of her ability to respond to novel equivalence relations.[70] She was taught the concept on a set of exemplars, but on the test sets she inferred logical equivalence of the stimuli *a* and *c*.[71]

Rio's performance is best explained as the result of non-associative learning giving rise to associative structures as well as of associative learning leading directly to the acquisition of propositional structures. She was trained on exemplars with trial

[70] Using differential reinforcement, Rio was taught the conditional relations aRb and bRc then aRa (symmetry), bRa (reflexivity), cRb (reflexivity), aRc (transitivity), cRa (reflexivity->equivalence) on problem sets 1-12 via trial-and-error and exclusion. She was then explicitly taught the conditional aRb and bRc relations on problem sets 13-30. The relations of symmetry, reflexivity, transitivity, and equivalence emerged without further training.

[71] Mandelbaum uses the phrase 'inferential promiscuity' to refer to trained relations that enter into further logical relations.

and error/positive reinforcement, which indicates that associative learning had taken place. But what she acquired seems to be more than an associative structure since she can apply it to novel stimuli sets. And she was also trained with an exclusion paradigm, which like fast-mapping is a case of one-shot learning that cannot be assimilated to associative learning even though she acquired associations between the members of the exemplar sets. But because of the productivity of what was learned, despite the mechanisms of learning, it is not possible to explain the contents of what she has learned as purely associatively structured.

What Rio learned (i.e. the concept of equivalence) might have been acquired via associative learning/conditioning, but following Mandelbaum I propose that what was acquired is not associatively structured. It could not be given its productivity, i.e. her ability to extend the concept of equivalence to novel relations. Associative structures are unable to account for the syntactic relations/symbolic reasoning demonstrated by Rio. It is possible, perhaps even probable, that associative learning on the problem sets gave rise to a conceptual understanding of the task at hand, which she was then able to apply to the testing problem sets. But a purely associative account of the learning, transitions, and structures/mechanisms cannot explain her successful performance on those testing sets. Rio has acquired concepts, not just responses to the particular stimuli she has been exposed to. So a pure associationism cannot account for the findings. Rio's ability to extend an equivalence concept to novel relations, albeit acquired via associative learning, shifts the burden of proof to the primate researchers proposing that all of the experimental results can be explained

by associatively structured mechanism simply because of the possibility of associative learning on some exemplars.

Works Cited

Albiach-Serrano, et al. “The Effect of Domestication and Ontogeny in Swine Cognition (*Sus scrofa scrofa* and *S. s. domestica*).” *Applied Animal Behaviour Science*, vol. 141, 2012, pp. 25-35.

Allen, Colin. “Transitive Inference in Animals: Reasoning or Conditioned Associations?.” *Rational Animals?*, edited by Susan Hurley and Matthew Nudds, Oxford University Press, 2006, pp. 175-186.

Anderson, Michael. “Neural Reuse: A Fundamental Organizational Principle of the Brain.” *Behavioral and Brain Sciences*, vol. 33, 2010, pp. 245-313.

Astington, Janet Wilde and Jennifer M. Jenkins. "A Longitudinal Study of the Relation Between Language and Theory-of-Mind Development." *Developmental Psychology*, vol. 35, no. 5, 1999, pp. 1311-1320.

Aust, Ulrike, et al. “Inferential Reasoning by Exclusion in Pigeons, Dogs, and Humans.” *Animal Cognition*, vol. 11, 2008, pp. 587-597.

Barad, Karen. “Agential Realism: Feminist Interventions in Understanding Scientific Practices.” *The Science Studies Reader*, edited by Marie Bagiote, Routledge, 1999, pp. 1-11.

Barad, Karen. *Meeting the Universe Halfway: Quantum physics and the Entanglement of Matter and Meaning*. Duke University Press, 2007.

Bermudez, Jose Luis. *Thinking without Words*. Oxford University Press, 2003.

Buxbaum, L.J. et al. “Relative Sparing of Object Recognition in Alexia-Prosopagnosia.” *Brain Cognition*, vol. 32, 1996, pp. 202-205.

Call, Josep. “Inferences by Exclusion in the Great Apes: The Effect of Age and Species.” *Animal Cognition*, vol. 9, 2006, pp. 393-403.

Call, Josep, et al. “’Unwilling’ Versus ‘Unable’: Chimpanzees’ Understanding of Human Intentional Action,” *Developmental Sciences*, vol. 7, no. 4, 2004, pp. 488-498.

Carey, Susan. “Beyond Fast Mapping.” *Language, Learning, and Development*, vol. 6, no. 3, 2010, pp. 184-205.

Carey, Susan. “The Child as Word Learner.” *Linguistic Theory and Psychological*

Reality, 2nd edition, edited by Morris Halle, Joan Bresnan, and George Miller, MIT Press, 1978.

Carey, Susan and Elsa Bartlett. "Acquiring a Single New Word." *Papers and reports on Child Language Development*, vol. 15, 1978, pp. 17–29

Carruthers, Peter. "Simulation and Self-knowledge: A Defense of Theory-Theory." *Theories of Theories of Mind*. Edited by Peter Carruthers and Peter K. Smith, Cambridge University Press, 1996, pp. 22-38.

Miozzo, Michele and Alfonso Caramazza, "Varieties of Pure Alexia— The Failure to Access Graphemic Representations," *Cognitive Neuropsychology*, vol. 15, 1998, pp. 203-238

Coltheart, Max. "Modularity and Cognition." *Trends in Cognitive Science*, vol. 3, no. 3, 1999, pp. 115-120.

Cosmides, Leda. "The Logic of Social Exchange: Has Natural Selection Shaped How Humans Reason? Studies with the Wason Selection Task." *Cognition*, vol. 31, no. 3, 1999, pp. 187-276.

Cosmides, Leda and John Tooby. "Evolutionary Psychology: A Primer." Center for Evolutionary Psychology, University of California Santa Barbara, 1997. (Available online at http://www.psych.ucsb.edu/research/cep/primer.html)

Dennett, Daniel. *Elbow Room: The Varieties of Free Will Worth Wanting*, MIT Press, Cambridge, Mass., 1984.

De Renzi, Ennio and Giuseppe Di Pellegrino. "Prosopagnosia and Alexia without Object Agnosia." *Cortex*, vol. 34, 1998, pp. 403-415.

Elsworthy, Charles. "Evolutionary Psychology: The Appropriate Disciplinary Link Between Evolutionary Theory and the Social Sciences." *The Darwinian Heritage and Sociobiology*, edited by Johan van der Dennen, David Smilie, and Daniel Wilson, Praeger Publishing, 1999, pp. 285-294.

Evans, Jonathon. "Matching Bias in Conditional Reasoning: Do We Inderstand it After 25 Years?" *Thinking and Reasoning*, vol. 4, no. 1, 1998, pp. 45-82.

Fitzpatrick, Simon. "The Primate Mindreading Controversy: A Case Study in Simplicity and Methodology in Animal Psychology." *The Philosophy of Animal Minds*, Cambridge University Press, 2009, pp. 258-277.

Flew, Antony. *A Dictionary of Philosophy: Revised Second Edition*. St. Martin's

Press, 1984.

Flombaum, Jonathon and Laurie Santos. "Rhesus Monkeys Attribute Perceptions to Others." *Current Biology*, vol. 15, 2005, pp. 447-452.

Fodor, Jerry. *Hume Variations*. Clarendon Press, 2003.

Fodor, Jerry. *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. MIT Press, 2001.

Fodor, Jerry. *The Modularity of Mind: An Essay on Faculty Psychology*. MIT Press, 1983.

Gallese, Vittorio and Alvin Goldman. "Mirror neurons and the Simulation Theory of Mind-reading." *Trends in Cognitive Sciences*, vol. 2, no. 12,1998, pp. 493-501.

Gallistel, Charles, et al. "The Learning Curve: Implications of a Quantitative Analysis." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 36, 2004, pp. 13124–13131.

Gallistel, Charles and Adam King. *Memory and the Computational Brain: Why Cognitive Science Will Transform Neuroscience*. Wiley-Blackwell, 2009.

Gardner, Richard Allen and Beatrix Gardner. "Teaching Sign Language to a Chimpanzee." *Science*, vol. 165, no. 3894, 1969, pp. 664-672.

Gardner, Richard Allen et al. (Editors). *Teaching Sign Language to Chimpanzees*, SUNY Press, 1989.

Gisiner, Robert and Ronald Schusterman. "Combinatorial Relationships Learned by a Language-Trained Sea Lion." *Marine Mammal Sensory Systems*, edited by J. Thomas et al., Plenum Press, New York, 1992.

Gordon, Robert. "'Radical' Simulationism." *Theories of Theories of Mind*, edited by Peter Carruthers and Peter K. Smith, Cambridge University Press, 1996, pp. 11-21.

Graham, George, "Behaviorism," *Stanford Encyclopedia of Philosophy*, (Fall 2016 edition), Edward N. Zalta (Ed.) forthcoming URL = <http://plato.stanford.edu/archives/fall2016/entries/behaviorism/>.

Griffin, Donald. *The Question of Animal Awareness*. Rockefeller University Press, 1976.

Griffin, Donald and Gayle B. Speck. "New Evidence of Animal Consciousness." *Animal Cognition*, vol. 7, 2004, pp. 5-18.

Hare, Brian, et al. "Chimpanzees know what conspecifics do and do not see." *Animal Behaviour*, vol. 59, 2000, pp. 771-785.

Hare, Brian, et al. "Do Chimpanzees Know What Conspecifics Know?" *Animal Behaviour*, vol. 61, 2001, pp. 139-151.

Hauser, Marc. "Our Chimpanzee Mind." *Nature*, vol. 437, 2005, pp. 60-63.

Hebb, Donald O. *A Textbook of Psychology*. W.B. Saunders Co., 1958.

Hermer, Linda and Elizabeth Spelke. "Modularity and Development: The Case of Spatial Reorientation." *Cognition*, vol. 61, 1996, pp. 195-232.

Hogan, Jerry. "The Structure Versus the Provenance of Behavior." *The Selection of Behavior: The Operant Behaviorism of B.F. Skinner*, edited by A. Charles Catania, Cambridge University Press, 1988, pp. 433-436.

Horowitz, Alexandra. "Theory of Mind in Dogs? Examining Method and Concept." *Learning and Behavior*, vol. 39, 2011, pp. 314-317.

Hume, David. *Treatise of Human Nature* Edited by L.A. Selby-Bigge, Clarendon Press, 1896.

Hume, David. *Enquiry Concerning Human Understanding*. 1748.

Kaminski, Juliane, et al. "Domestic Goats, *Capra hircus*, Follow Gaze Direction and Use Social Cues in an Object Choice Task." *Animal Behaviour*, vol. 69, 2004, pp. 11-18

Kastak, David and Ronald Schusterman. "Comparative Cognition in Marine Mammals: A Clarification on Match-to-Sample Tests." *Marine Mammal* Science, vol. 8, no. 4, 1992, pp. 414-417.

Köhler, Wolfgang, *The Mentality of Apes*. 2nd edition, Springer, 1921. Translated by Ella Winter. 1925. Liveright, 1976.

Locke, John. *Essay Concerning Human Understanding*. 4th edition, 1700.

Maginnity, Michelle. "Perspective Taking and Knowledge Attribution in the

Domestic Dog (*Canis Familiaris*): A Canine Theory of Mind?" Thesis submitted to the University of Canterbury, 2007.

Mandelbaum, Eric. "Attitude, Inference, Association: On the Propositional Structure of Implicit Bias." *Nous*, online first, 2014, pp. 1-30.

Mandelbaum, Eric. "Associationist Theories of Thought", *The Stanford Encyclopedia of Philosophy* (Summer 2016 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2016/entries/associationist-thought/>.

Markson, Lori, and Paul Bloom. "Evidence against a Dedicated System for Word Learning in Children." *Nature*, vol. 385, no. 6619, 1997, pp. 813-815.

Mitchell, Chris, et al. "The Propositional Nature of Human Associative Learning." *Behavioral and Brain Sciences*, vol. 32, 2009, pp. 183-246.

Morgan, Conway Lloyd. *Introduction to Comparative Cognition*. W. Scott Limited, 1894.

Morgan, Conway Lloyd. *Introduction to Comparative Cognition*. 2nd edition, W. Scott Limited, 1903.

Neisser, Ulrich. *Cognitive Psychology*. Appleton-Century-Crofts, 1967.

O'Donnell, Jennifer and Kathryn Saunders. "Equivalence Relations in Individuals with Language Limitations and Mental retardation." *Journal of the Experimental Analysis of Behavior,* vol. 80, no, 1, 2003, pp. 131-157.

Okrent, Mark. *Rational Animals: The Teleological Roots of Intentionality*. Ohio University Press, 2007.

Onishi, Kristine and Renee Baillargeon. "Do 15-month-old Infants Understand False Beliefs?" Science, vol. 308, 2005, pp. 255-258.

Pavlov, Ivan. *Conditioned Reflexes; An Investigation of the Physiological activity of the Cerebral Cortex***.** Oxford University Press, 1927**.**

Pavlov, Ivan. *Lectures on Conditioned Reflexes*. Liveright, 1928.

Penn, Derek, et al. "Darwin's Mistake: Explaining the Discontinuity Between Human and Nonhuman Minds." *Behavioral and Brain Sciences*, vol. 31, 2008, pp. 109-178.

Piaget, Jean. *The Construction of Reality in the Child*. New York Basic Books, 1954.

Pilley, John and Alliston Reid. “Border Collie Comprehends Object Names as Verbal Referents.” *Behavioural Processes*, vol. 86, 2011, pp. 184-195.

Pinker, Steven. *The Language Instinct: How the Mind Creates Language*. 1994. Harper Perennial, 2007.

Povinelli, Daniel, et al. “Inferences About Guessing and Knowing by Chimpanzees (*Pan troglodytes*).” *Journal of Comparative Psychology*, vol. 104, no. 3, 1990, pp. 203-210.

Povinelli, Daniel and Timothy Eddy. “What Young Chimpanzees Know About Seeing.” *Monographs of the Society for Research in Child Development*, vol. 6, 1996, pp. 1-152.

Povinelli, Daniel and Steve Giambrone. “Reasoning About Beliefs: A Human Specialization?” *Child Development*, vol. 72, 2001, pp. 691-695.

Premack, Anne. *Why Chimps Can* Read. Harper and Row, 1976.

Premack, David and Guy Woodruff. “Does the Chimpanzee Have a Theory of Mind?” *Behavioral and Brain Sciences*, vol. 1, no. 4, 1978, pp. 515-626.

Rescorla, Robert A., and Allan R. Wagner. "A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement." *Classical conditioning II: Current research and theory*, 1972, pp. 64-99.

Rumiati, Raffealla and Glyn Humprheys. “Visual Object Agnosia without Alexia-Prosopagnosia.” *Visual Cognition*, vol. 4, 1997, pp. 207-217.

Samuels, Richard. “Massively Modular Minds: Evolutionary Psychology and Cognitive Architectures.” *Evolution and the Human Mind: Modularity, Language, and Meta-cognition*, edited by Peter Carruthers and Andrew Chamberlain, Cambridge University Press, 2000, pp. 13-46.

Samuels, Richard and Stephen Stich. “Rationality and Psychology.” *The Oxford Handbook of Rationality*, edited by Alfred Mele and Piers Rawling, Oxford University Press, 2004, pp. 279-300.

Santos, Laurie, et al. "The Evolution of Human Mindreading: How Non-human Primates Can Inform Social Cognitive Neuroscience." *Evolutionary Cognitive Neuroscience*, edited by Steven Platek et al., MIT Press, 2006.

Santos, Laurie R. and Marc Hauser. "How Monkeys See the Eyes: Cotton-top

Tamarins' Reaction to Changes in Visual Attention and Action." *Animal Cognition*, vol. 2, 1999, pp. 131-139.

Santos, Laurie, et al. "Rhesus Monkeys, *Macaca Mulatta*, Know What Others can and Cannot Hear." *Animal Behavior*, vol. 71, 2006, pp. 1175-1181.

Schusterman, Ronald and Robert Gisiner. "Pinnipeds, Porpoises, and Parsimony: Animal Language Research Viewed from a Bottom-up Perspective." *Anthropomorphism, Anecdotes, and Animals: The Emperor's New Clothes?*, edited by Robert Mitchell et al. SUNY Press, 1997, pp. 370-382.

Schusterman, Ronald and David Kastak. "A California Sea Lion (*Zalophus Californianus*) is Capable of Forming Equivalence Relations." *The Psychological Record*, vol. 43, 1993, pp. 823-839.

Schusterman, Ronald and David Kastak. "There is No Substitute for an Experimental Analysis of Marine Mammal Cognition." *Marine Mammal Science*, vol. 11, no. 2, 1995, pp. 263-267.

Sextus Empiricus. *Outlines of Pyrrhonism, Vol. 1*, translated by R.G. Bury, Harvard University Press, 1933.

Sidman, Murray. "Reading and Auditory-Visual Equivalence." *Journal of Speech and Hearing Research*, vol. 4, no. 1, 1971, pp. 5-13.

Sidman, Murray and William Tailby. "Conditional Discrimination vs. Matching to Sample: An Expansion of the Testing Paradigm." *Journal of Experimental Analysis of Behavior*, vol. 37, no. 1, 1982, pp. 5-22.

Skinner, Burrhus F. *The Behavior of Organisms: An Experimental Analysis*, D. Appleton and Company, 1938.

Skinner, Burrhus F. *Science and Human Behavior*, Macmillan, 1953

Sober, Elliot. "Parsimony and Models of Animal Minds." *The Philosophy of Animal Minds*, edited by Robert Lurz, Cambridge University Press, 2009, pp. 237-257.

Sober, Elliott. "The Principle of Parsimony." *British Journal for the Philosophy of Science*, vol. 32, no. 2, 1981, pp. 145-156.

Thomas, Roger K. "Lloyd Morgan's Canon." *Comparative Psychology: A*

Handbook, edited by G. Greenberg and M.M. Haraway, Garland Publishing Co., 1998, pp. 156-163.

Thomas, Roger K. "Lloyd Morgan's Canon: A History of Misrepresentation." Originally published in *History and Theory of Psychology ePrint Archive*, York University. University of Georgia, 2001, https://faculty.franklin.uga.edu/rkthomas/sites/faculty.franklin.uga.edu.rkthomas/files/MCPrintOptimal.pdf . Accessed 15 October 2016.

Thorndike, Edward L. *Animal Intelligence*. Macmillan, 1911.

Toates, Frederick. "A Model of the Hierarchy of Behaviour, Cognition, and Consciousness." *Consciousness and Cognition*, vol. 15, no. 1, pp. 75-118.

Tolman, Edward. "Cognitive Maps in Rats and Men." *The Psychological* Review, vol. 55, no. 4, 1948, pp. 189-208.

Tomasello, Michael and Josep Call. *Primate Cognition*. Oxford University Press, 1997.

Tomasello, Michael, et al. "Chimpanzees Understand Psychological States— the Question is Which Ones and to What Extent." *TRENDS in Cognitive Sciences*, vol. 7, no. 4, 2004, pp. 153-156.

Trivers, Robert. "The Evolution of Reciprocal Altruism." *The Quarterly Review of Biology*, vol. 46, no. 1, 1971, pp. 35-57.

Wason, Peter Cathcart. "Reasoning." *New Horizons in Psychology Volume 1*, edited by B.M. Foster, Penguin, 1966, pp. 135-151.

Wason, Peter Cathcart. "Reasoning about a Rule," *The Quarterly Journal of Experimental Psychology*, vol. 20, no. 3, 1968, pp. 273-281.

Wason, Peter Cathcart and Philip Nicholas Johnson-Laird. *Psychology of Reasoning: Structure and Content*, Harvard University Press, 1972.

Watanabe, Shigeru and Ludwig Huber. "Animal Logics: Decisions in the Absence of Human Language." *Animal Cognition*, vol. 9, 2006, pp. 235-245.

Watson, John B. "Psychology as the Behaviorist Views it." *Psychological Review*, vol. 20, 1913, pp. 158-177.

Watson, J.B. *Behavior: An Introduction to Comparative Psychology.* Henry Holt, 1994.

Welsh, Matthew Brian. "Of Parrots and Parsimony: Reconsidering Morgan's Canon." *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, edited by S. Ohlsson and R. Catrambone, 2010, pp. 1798-1803.

Wilson, Edward O. *Sociobiology: The New Synthesis*. Harvard University Press, 1975.

Wooldridge, Dean. *Mechanical Man: The Physical Basis of Intelligent Life*. McGraw Hill, 1968.

Yerkes, Robert Mearns. *Almost Human*. J. Cape, 1925.