Cognitive Critique

A FUNCTIONALIST VIEW OF METACOGNITION IN ANIMALS

NISHEETH SRIVASTAVA

Department of Computer Science
University of Minnesota, Minneapolis, Minnesota

EMAIL: nsriva@cs.umn.edu

C. WADE SAVAGE

Center for Cognitive Sciences
University of Minnesota, Minneapolis, Minnesota

EMAIL: savag001@umn.edu

Accepted November 1, 2012

KEYWORDS

metacognition, animal cognition, theory of mind, non-linguistic thought, structuralism

ABSTRACT

We present a structuralist analysis of the current state of cognitive science research into the phenomenon of metacognition. We begin from the assumption that cognitive intelligence is an organ just like any other biological organ, with the defined function of allowing intelligent entities to maintain homeostasis with a changing but predictable environment. This understanding leads to the conclusion that human cognition is functionally derived to generate accurate preference relations about the world. Examining empirical evidence from recent research using definitions of metacognition and theory of mind that emerge from our functionalist understanding of cognition, we conclude decisively in favor of the existence of metacognition in non-human animals.

BACKGROUND

Metacognition is generally considered to be cognition about cognitive processes, namely second order cognition, and has heretofore been thought to be a uniquely human property. Partly because of the importance of this concept in human species-chauvinism, research into non-human animal meta-cognition has had to contend with philosophical controversies. While philosophical controversies are generally quite harmless, in this particular domain erroneous philosophical arguments can and indeed do affect experimental setup and resulting interpretation significantly. Where to one observer, an animal pointing to a toy and making urgent noises unequivocally represents the thought *I want that*, another observer claims that such an interpretation is biased and that in the absence of language, ascribing thoughts to animals is wrong. These debates have existed in animal cognition research since the very inception of the field, and they arise because of the ambiguous definitions of the fundamental concepts under study. Unless concepts like thought, belief, intention, etc. are unambiguously defined, there can be no hope for philosophical agreement in animal cognition research.

It is therefore important to establish a strong epistemological basis for discussions on animal cognition, one that is isomorphic with the reasonable epistemologies proposed before, and that also invalidates the erroneous claims made previously in the literature. The first part of this paper uses evolutionary arguments to propose a sound epistemological basis for discussing animal cognition concepts. The second part demonstrates the value of the approach by using it to explain and resolve existing debates on the nature of animal metacognition using empirical evidence. We conclude with some remarks on the salience of animal metacognition research to our understanding of human metacognition and on the knottier question of the nature of *consciousness*.

While one stream of research in metacognition examines the ability of animals to reliably attribute to others and draw accurate inferences from them, another considers the ability to assign uncertainty and other properties to their own beliefs. Current research reviewed in Penn and Povinelli (2007) and Call and Tomasello (2008) exemplifies advances made in the first domain, generally identified as the theory of mind (ToM) research community. Good examples of research conducted in the second domain, typically known as uncertainty monitoring, include Shettleworth and Sutton (2006) and Call and Carpenter (2001).

This paper briefly examines the current state of affairs in both these domains and demonstrates that some of the controversy in ToM research (especially the epistemological divide between the Povinelli and Tomasello groups) is unfounded. It further addresses the issue of ecological validity which has emerged as a point of contention concerning some of the more sophisticated experimental protocols proposed for testing ToM.

THE EVOLUTIONARY ORIGIN OF INTELLIGENCE

It is somewhat surprising, given the universally acknowledged validity of evolution driven by natural selection as the generative mechanism for all biological phenomena, to find that it is even possible to hold the view that non-human animals are somehow categorically disqualified from possessing sophisticated cognitive concepts such as beliefs, thoughts and self-awareness. While some degree of anthropomorphic prejudice is understandable, and the cultural dominance of *homo sapiens* is absolute beyond question, it is still unscientific to assume that cognitive capabilities have somehow arisen in human beings through mechanisms unlike those that have brought various other species to their respective relational equilibria with the biosphere. We find it difficult to entertain a view other than the one suggesting that cognitive capacities have arisen in multiple species through natural selection using the response to stimulus as the fitness criterion. This view suggests that many species will manifest cognitive capabilities similar in several respects to human beings, in some correspondence with their shared evolutionary ancestry.

It is therefore unsurprising that in higher primates we find evidence of human-like intelligence, a fact that, incredibly, some observers have criticized as an anthropomorphic bias. The biological and evolutionary basis of the origin of cognitive concepts also brings into question definitions based on linguistic and other arbitrary bases. Some organisms, in the course of random speciation, developed the ability to respond to their environment in ways that made their survival more likely. This ability to select a subset from a number of possible outcomes can be understood more generally as a preference relation developed by the organism. For several organisms, this level of informational processing remained their cognitive zenith. Contemporary examples of such creatures include almost all

the lower phylae of the animal kingdom. In time, the descendants of organisms that had developed the ability to form simple preference relations further developed the ability to form increasingly sophisticated preference relations about objects behaving in increasingly complex ways. Some such organisms, for various evolutionary reasons, lived in large herds populated by other members of the same species. Their immediate environment was therefore generally comprised of organisms similar to themselves, whose behavior it was in their evolutionary interests to predict. Thus, for social animals, the desire to predict outcomes in the immediate environment proved to be evolutionarily equivalent to developing what we now understand to be ToM.

This conjecture concerning the origin of ToM is eminently testable. For instance, it offers a simple explanation for the existence of sophisticated social behavior in organisms that appear to have extremely rudimentary cognitive apparatus, e.g., bees, ants. In such cases, the immediate environment of the organism's ancestors was populated with a very high density of similar individuals, making the ability to predict other's behaviors a strongly selected for trait. We further hypothesize that the degree to which ToM is observed in various organisms will be positively correlated with the density of the social interactions in their species. Empirical observations in this regard should prove to be instructive.

Similarly, for creatures whose environments were susceptible to rapid changes, it was evolutionarily beneficial to evolve a meta-level of certainty with respect to pre-existing preference relations, thereby leading to the selection of self-reflective metacognitive abilities. As in the previous case, we make a testable prediction with respect to our conjecture. We hypothesize that organisms whose natural mode of existence requires a greater density of selection decisions will more likely possess self-awareness.

EVOLUTION AND STRUCTURALIST DEFINITIONS

If we agree on the evolutionary origin of intelligence, it becomes evident that epistemological questions concerning cognition cannot be answered completely without reference to their generative process. Since the process of evolutionary selection is predicated on the development of preference relations in organisms, and since preference relations are measurable in terms of statistical observa-

tions of behavior, a definitional framework of cognitive concepts is possible by viewing them as evolutionary artifacts arising from the development of increasingly sophisticated preference relations. It may be argued that other definitions of cognitive concepts cannot be disregarded, and that the framework presented here is too simplistic to account for the complexity of existing cognitive phenomena. However, since our definitions proceed entirely from evolutionary theory and are completely structural in the sense that they are formulated independent of the semantic objects they seek to define, we argue that any alternative definitions of cognitive concepts will either be isomorphic with or special cases of our own framework, or will be incorrect. In particular, instead of examining the vexatious questions, *Does a dog have a thought in the way a man has a thought? How are they the same? How are they different?* we instead ask, *Is the evolutionary process that allows a dog to interact with its environment the same as the evolutionary process that allows a man to deal with his environment?* Since this answer is trivially affirmative, we are now freed to inquire into the similarities and dissimilarities between the man's cognitive processes and the dog's, without entering into the semantic morass that arises in wondering what it is for a man to have a thought. By invoking evolutionary arguments and removing man's special place in the scale of cognition, we clarify our understanding of the processes involved in animal cognition without being sidetracked by semantic disputes. Metacognition or self-reflective consciousness has been central in the distinction between humans and other species to the extent that it is generally used as a definition for sapience, which in turn is used as a defining characteristic for our species. Attributing metacognition to non-human animals may thus become an emotive issue for observers who prefer to consider humans as distinct from other species. Given the potential for controversy, it is essential that an articulation of terminology and concepts be made before an evaluation of empirical evidence is attempted. This is precisely what we endeavor to do in this section.

Our definitional framework of cognitive concepts avoids the behaviorist-mentalist debates over their epistemology by defining them in an ontologically agnostic manner. For example, from a structuralist point of view, a subject's observation that a certain organism prefers a particular outcome over other outcomes is sufficient for the subject to ascribe belief to the organism. Questioning whether the organism actually possesses this belief thus becomes

philosophically identical to asking, *Do photons exist?* Just as physicists find it simpler to describe some aspects of the behavior of light by positing the existence of massless particles, subjects find it easier to describe an entity's behavior by imbuing it with a preference relation, thereby narrowing their own hypothesis space for predicting its behavior.

We show that our definitions allow us to evade the ideological battles prevalent in animal minds research and make clear logical determinations of the validity or invalidity of experimental observations with respect to animal metacognition.

Metacognition is consensually defined as cognition of cognition. Multiple definitions of cognition abound in the literature, but few are rigorously grounded. For this paper, we suggest a general definition of cognition as a process by which an organism changes its preferences with respect to the objects the process is operating upon. Thus, metacognition may be defined to be a process by which an agent changes its preferences with respect to other agent-specific preference relations. In the interest of completeness, preference relations may be defined as subjectively held judgments of value over a set of feasible outcomes.

SELF-AWARENESS

Mental processes with respect to other agent-specific preference relations are further subdivided based on the identity of the agent (the one that holds preference relations) and the meta-agent (the one that processes and assigns values to these preference relations). In the case where the identities of the agent and the meta-agent are the same, metacognition is equivalent to self-awareness and subjective uncertainty manipulation. For example, the questions *Am I correct to hold the belief X?*, *Am I more sure about belief X than belief Y?*, *Do I possess the property P because I hold the belief X?* are all examples of self-referential metacognition and, therefore, self-awareness. To formalize this notion of self-awareness, we require two further definitions. First, we define beliefs to be isomorphic to preference relations. Second, we define existential properties as properties that agents self-regard (independent of duration) as components of their identity.

With these terms defined, we define self-awareness as a property characterized by the ability to formulate beliefs expressible as statements of the form, *Existential property P holds (to a certain degree) because I hold (to a certain degree) the belief X as opposed*

to other beliefs. That is, we define self-awareness as the ability to differentiably identify beliefs that one possesses.

THEORY OF MIND

In the case of metacognition where the identity of the agent is not the same as the identity of the meta-agent, we impute to the meta-cognitive meta-agent the property ToM. Using the terminology we have developed, we define ToM as a property characterized by the ability to formulate beliefs expressible as statements of the form *Property P holds (to a certain degree) because agent A holds (to a certain degree) the belief X as opposed to other beliefs*. That is, we define ToM as the ability to differentiably identify beliefs that (we infer) other agents possess.

Premack and Woodruff (1978), in their seminal work on the subject, define ToM as, *the ability to predict and explain behavior by attributing mental states*. Understanding predicting and explaining as important to forming preferences, identifying mental states with beliefs and taking behavior to be the specific implementation of beliefs held, it is clear that our definition subsumes the original. Furthermore, as we shall below, our definition also allows us to differentiate between two competing definitions of ToM that have recently generated much polemic in the animal cognition research community.

EMPIRICAL OBSERVATIONS

Empirical research into animal cognition presents a plethora of possibilities for analysis. We choose to focus on a small subset of recent findings that best conform to our structuralist definitions of metacognition, and that have occasioned theoretical disputes with respect to their interpretation. We first consider empirical evidence concerning self-reflective abilities in non-human animals.

EVIDENCE FOR SELF-AWARENESS

An area of much attention in self-awareness testing is the mirror self-recognition test (MSR), in which animals are surreptitiously marked with a colored spot and then allowed access to a mirror. Greater than random frequency of touching the colored region is considered to be evidence of self-recognition. Higher primates, dolphins, and elephants have passed the MSR test (Andrews 2008).

That self-recognition is sufficient evidence for ascribing self-awareness to a creature is a disputable claim. It is possible to formulate the animal's attempts to touch the colored spot as the following statement: *Existential property P holds because I hold belief X*, where P is the property of having a spot and X is the belief *the reflection in the mirror is me*. This formulation, however, rests on the animal's ability to accurately gauge the function of the mirror and so may be challenged on multiple grounds. For example, a behaviorist interpretation would suggest that the animal finds that it can reach the spot in the mirror either by touching the mirror or by moving its limbs in ways that correspond with the spot on its own body. Thus, the behaviorist may persuasively argue that the greater than random frequency of touching the colored region might occur through simple goal-directed exploratory behavior without self-awareness.

A more suitable test for self-awareness is uncertainty monitoring, where creatures' confidence in their beliefs are tested. The standard protocol for most such experiments is as follows: the subject is trained using rewards for correct choices and no rewards for incorrect choices. Then, a third *bailout* choice is introduced with a smaller reward than the one where the subject performed the task correctly. The stimuli used in the test vary in ambiguity and therefore can cause difficulty in judgment. If subjects learn to use bailout options in cases where stimuli are ambiguous, we can conclude that they are aware of the uncertainty of their *beliefs*.

Unlike in the case of the MSR, no other explanation appears to be admissible here. Successful behavior in this instance may be formulated as a preference relation of the form *Existential property P holds because I hold belief X*, where P is the agent's sense of certainty of solving the problem and X is the agent's preference relation with respect to the outcomes of the experimental test. Note that unlike in the case of the mirror test, belief X in this instance does not depend on a particular interpretation of the agent's understanding of its environment but constitutes a domain-independent appraisal of its preference for each one of the possible outcomes of the test. Thus, any other explanation of success in this domain would be structurally indistinguishable from the one above that ascribes to the agent in question the ability to form a meta-level preference relation over its lower level preference relations. By the definition we have proposed, such an agent may be said to possess self-awareness. Call and Tomasello (2008) report that both great apes and human infants learn to use the bailout choice to maximize their potential reward. Similar behavior has been seen in dolphins,

rhesus monkeys (Andrews 2008) and rats (Foote and Crystal 2007). The evidence in favor of non-human self-awareness, viewed in the sense of our definition, thus appears to be incontrovertible.

EVIDENCE FOR THEORY OF MIND

The fundamental difficulty in studying ToM is epistemological and is described by Dennett as being related to the problem of other minds. His proposed solution, the intentional stance (Dennett 2009), has not been unequivocally accepted by the experimental research community. It is entirely possible to find, as indeed we do, the same experimental results being interpreted along both behaviorist and mentalist lines in the ToM research community. A case in point is the study conducted on chimpanzees in Hare et al. (2001), which is presented as evidence in favor of the presence of ToM in chimpanzees by Call and Tomasello (2008) while being called ineffectual in principle by Penn and Povinelli (2007).

The experiment in question is designed as follows. Dyads of chimpanzees, one dominant (called Alice) and one non-dominant (called Bob), are selected by ethological observation. Both chimpanzees are placed in separate boxes within each others' sight and food is placed in a third compartment (one out of multiple identical compartments) that both have access to. Shades blocking Alice's view of the food compartment may be drawn and are visible to Bob just as well. The experiment consists of three stages. In the first stage, Alice is not present when food is placed in a food compartment and brought into the room. In the second stage, Alice is present when the food is placed in a food compartment, as is Bob. In the third stage, Alice is present when the food is placed in a food compartment, then removed and replaced in another compartment, whereupon she is retrieved to her original position. Call and Tomasello (2008) found that Bob could distinguish between the first two stages, i.e., he could know when Alice knew or did not know about the food being placed in a compartment. However, Bob was unable to distinguish between the second and the third stages, i.e., he did not ascribe a false belief to Alice. This causes Call and Tomasello (2008) to suggest that although chimpanzees possess ToM in the sense of being able to predict the beliefs of others, they do not possess the ability to judge others' beliefs counterfactually.

Let us examine the evidence obtained through the perspective of our definitions. Bob's distinction between the first two stages of the experiment may be formulated as his subscription to a statement

of the form, *Property P holds because Alice holds belief X*, where P is the relative safety of Bob taking the food and X is the belief that food has been placed in some compartment. In this case, while the formalism satisfies our definition, the belief X is not domain-independent and hence, our interpretation is not yet shown to be structurally isomorphic with any others that may be produced. The fact that the test is conducted in a competitive scenario concerning food has caused the Povinelli group to raise objections concerning the ecological validity of the experimental setup. Their basic argument is that the *naturalness* of the environment allows Bob to differentiate between the first two stages of the experiment.

Thus, Penn and Povinelli (2007) suggest that the test outlined above is ineffectual in detecting the existence of ToM, since the chimpanzees differentiation between the first and second stages of the experiment can be explained by behavioral cues. They suggest that since the domain under consideration is one of food competition, the ape Bob may conceivably possess a behavioral cue of the form, *Do not take food if Alice's eyes are facing food*. They propose an alternative test that would eliminate the potential for such behaviorist explanations in the form of an opaque vizor test, in which the subject ape is introduced to two vizors of different colors – one red and one green. The red vizor's eye-holes are painted over making it impossible to see through them. Then, the ape is made to beg for food from an experimenter who is wearing either the red or the green vizor. If the ape does not beg with as much intensity when confronted with the experimenter wearing the opaque vizor, he can be said to have ToM since no behaviorist cue could allow the ape to associate the concept of *not seeing* with the red vizor. Penn and Povinelli (2007) further claim that experiments conducted by the Povinelli group and reported in an, thus far, unpublished study show that apes fail this test, although a modification of the test is passed by 15 month old infants.

While their methodology has been criticized (Andrews 2009) as being susceptible to their own behaviorist criticism, a far more serious objection is raised on the grounds of the ecological validity of the experimental protocol. Remarkably, both groups accuse the other of failing this control for opposite reasons. The Povinelli group, as stated earlier, suggests that tasks that require food competition are likely to invoke behavioral cues in subjects. The Tomasello group suggests that tasks such as the opaque vizor test force the subjects into making decisions in ecologically unnatural environments, which makes negative conclusions drawn from such tests

unacceptable. Both complaints appear to have some merit, which leaves us at an impasse, wherein both positive results obtained in ecologically natural settings and negative results obtained in ecologically unnatural settings are disallowed. Resolution may perhaps be achieved through the following thought experiment.

Assume that the opaque vizor test is conducted on red-green color-blind adults. It is highly unlikely that any would pass the test as it currently stands. This would lead Penn and Povinelli (2007) to claim that color-blind adults do not demonstrate evidence of ToM, which is an unreasonable position. This suggests that all formulations of the opaque vizor test, non-verbal false belief tests in general, require cognitive capacities in addition to ToM, namely the ability to employ concepts that allow for differentiation between true and false beliefs as defined in the test. For instance, in Hare et al. (2001), the fact that Bob could not differentiate the third stage from the first one does not necessarily imply an inability to reason counterfactually. It is plausible that for Bob, any object placed out of sight in a downward direction is *on the ground*, and the fact that there are multiple compartments that are visuo-spatially exclusive on the floor does not occur to him in this context. Thus, his inability to judge what Alice would see does not imply a deficiency in his inference about Alice's actions. It is equally plausible that his inability to judge arises from his inability to *see* in a way that is natural to human experimenters.

While the Povinelli camp suggests that ecologically natural environments do not allow experimenters to test subjects' ability to *abstract out* from the problem domain and demonstrate true intelligence, our thought experiment suggest that the ability of humans to abstract out concepts like color, shape, etc are also governed by the linguistic ecology in which we are typically immersed. Thus, we conclude that such arguments against animal intelligence could also be directed against human intelligence, which is certainly inappropriate. Thus, the Povinelli argument against drawing positive results from ecologically natural test settings is refuted on the grounds that accepting it would require us to be skeptical of humans' possession of ToM. Remarkably, we find that we have managed to construct what closely resembles Davidson's infamous argument (Davidson 2001) against animal cognition (i.e., animals do not possess a linguistic ecology necessary for formulating thought) as a supplementary one in favor of animal cognition!

To summarize, sufficient evidence exists in favor of the view that some non-human animals, e.g., chimpanzees, possess the ability to infer what other agents know or do not know and to direct their own actions accordingly. The question whether animals have the ability to reason counterfactually is complicated by the necessity of agreement between experimenter and test subject regarding the concept space of the experimental procedure. Hence, negative results on the matter cannot be taken to be conclusive. More importantly, following our definitions, counterfactual reasoning capability is distinct from possession of ToM. With the caveat that ecological naturalness is assumed to be conceptually benign in judging functional performance, we may conclude that the chimpanzees in Hare et al. (2001) unambiguously satisfy the criteria for possession of ToM according to our definitions. With respect to the caveat we are required to assume, the simple leveling argument arising from our thought experiment above precludes the validity of any meaningful results being obtained from experimental setups deemed ecologically unnatural. Since only results obtained in ecologically natural settings are acceptable, our caveat becomes an extremely reasonable assumption. Thus, we believe the question whether some non-humans possess ToM is answered definitively in the affirmative.

CONCLUSION

As with any other philosophical debate concerning the validity of empirical observations, finding a reasonable framework for defining controversial terms is the largest part of the battle. In the domain of animal cognition, the philosophical debate has tended to revolve around the ontological significance of various cognitive concepts. Observing some similarities between this debate and the one engendered by the quest for the foundations of mathematics in the early part of the 20th century (see, e.g., Hellman 2005), we have endeavored to find a resolution analogous to the one discovered by the mathematical structuralists, that is, to define concepts using relations between observable objects of interest. This insight, combined with evolutionary arguments already extant (in, e.g., Dennett 1994), has allowed us to develop a novel definitional framework for understanding metacognition. While we believe such a structuralist approach to defining psychological concepts in evolutionary terms is generalizable (and resembles Anderson's [1990] rational analysis program in some ways), we have limited ourselves in this paper to

analyzing empirical research into non-human metacognition.

Our approach allows us to differentiate two distinct lines of inquiry into animal metacognition. The first, self-awareness, appears to be conclusively established, not just for primates, but also for several other mammals. The second, ToM, involving as it does some tricky manipulation around the problem of other minds, is not yet clearly settled. However, experimental results, interpreted in our structural framework, appear to leave little doubt that chimpanzees, at least, possess ToM. We further see how disputes concerning these results have arisen primarily out of a regrettable conflation of ToM with various linguistic capabilities through insistence on false belief testing.

Throughout this exegesis, the structuralist approach to defining concepts clarifies perspective with respect to the knotty philosophical debates that populate this area of the literature. We suggest, therefore, that it might also be useful in analyzing other disputes in the field.

ACKNOWLEDGMENTS

NS acknowledges several discussions with Iman Chahine as contributory to the germination and completion of this work.

REFERENCES

Anderson JR (1990) *The adaptive character of thought*. Erlbaum, Hillsdale, NJ

Andrews K (2008) Animal cognition. In: Zalta EN (ed) *The Stanford encyclopedia of philosophy* (Winter 2008 Edition) http://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi

Andrews K (2009) Politics or metaphysics? On attributing psychological properties to animals. Biol and Philos, 24:51-63

Call J, Carpenter M (2001) Do apes and children know what they have seen? Anim Cogn 4:207-220

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

Davidson D (2001) Rational animals. *Subjective, intersubjective, objective*. Clarendon Press, Oxford, UK, pp 95-106

Dennett D (1994) The role of language in intelligence. In: Khalfa J (ed) *What is intelligence? The Darwin College lectures*. Cambridge University Press, Cambridge, UK, pp 161-178

Dennett D (2009) Intentional systems theory. In: McLaughlin B et al. (eds) *The Oxford handbook of philosophy of mind*. Oxford University Press, Oxford, UK, pp 339-350

Foote AL, Crystal JD (2007) Metacognition in the rat. Curr Biol 17:551-555

Hare B, Call J, Tomasello M (2001) Do chimpanzees know what conspecifics know and do not know? Anim Behav 61:139-151

Hellman G (2005) Structuralism. In: Shapiro S (ed) *The Oxford handbook of philosophy of mathematics and logic*. Oxford University Press, Oxford, UK, pp 536-562

Penn DC, Povinelli DJ (2007) On the lack of evidence that non-human animals possess anything remotely resembling a theory of mind. Philos Trans R Soc Lond B Biol Sci 362:731-744

Premack D, Woodruff G (1978) Does the chimpanzee have a theory of mind? Behav Brain Sci 1:515-526

Shettleworth SJ, Sutton JE (2006) Do animals know what they know? In: Hurley S, Nudds M (eds) *Rational animals?* Oxford University Press, New York, NY, pp 235-246