Snipping or Editing? Parsimony in the Chimpanzee Mindreading Debate
Kristin Andrews

Metascience
Symposium on Elliott Sober's book *Ockham's Razors*

Advice about how to move forward on the mindreading debate, particularly when it comes to overcoming the logical problem, is much needed in comparative psychology. In chapter 4 Sober takes on the task by suggesting how we might uncover the mechanism that mediates between the environmental stimuli that is visible to all, and chimpanzee social behavior.

A little background. The starting assumption in this debate is that humans reason about others' mental states when predicting behavior, and the question is whether chimpanzees also reason about others' mental states when successfully predicting behavior, or whether they reason about others' behavioral tendencies given the stimuli. In other words, the question is whether chimpanzees are mindreaders or behavior-readers.

Sober presents this puzzle as a case of blackbox inference, according to which the competing hypotheses are represented as flow charts that connect the stimuli to the behavior via different intervening variables. The problem is set up in the context of Brian Hare, Josep Call, and Michael Tomasello's food competition studies, in which they offered subordinate chimpanzee subjects a chance to compete for food with a dominant chimpanzee (Hare et al. 2000, 2001). The subordinate and dominant chimpanzees were released into a room via two different doors, and they could see one another (or the closed door) across the room. Because dominant chimpanzees do not tolerate subordinate chimpanzees taking their food, the subordinate wouldn't take food that he expected the dominant would try to eat. In some experimental conditions a piece of food was hidden from the dominant, but it was visible to the subordinate. In such cases the subordinate chimpanzees avoided food that the dominants could see, or saw being hidden, and would seek out food that the dominant did not see. Hare and colleagues conclude that the chimpanzees understand what others can see. Advocates of chimpanzee mindreading appeal to parsimony insofar as the mindreading hypothesis unifies a number of different behaviors, rather than offering different rules for each (Tomasello and Call 2006, and Sober identifies Whiten 1996).

In their own appeal to parsimony, Daniel Povinelli and Jennifer Vonk disagree that this study provides evidence of chimpanzee mindreading, because they think that the behavior is consistent with an alternative behavior-reading hypothesis. Their version of parsimony takes the form of the razor of silence; we need not appeal to mental state attribution in order to explain the behavior, but those who appeal to mental state attribution also has to appeal to behavior rules. Thus, Sober suggests, the behavior-reading hypothesis permits snipping an intervening variable.

Sober states the two hypotheses in terms of the dominant chimpanzee (D) and subordinate chimpanzee (S):

(M) D's behavior → S's beliefs about D's behavior → S's beliefs about D's mind → S's behavior

(B) D's behavior → S's beliefs about D's behavior → S's behavior (213)

Notice that in the mindreading hypothesis (M) there are two intervening variables, while in the behavior-reading hypothesis (B) there is only one. B snips the second intervening variable M postulates, and that is where the claim of parsimony comes from.

Povinelli and Vonk's argument in favor of something like B stems from their worry about what they call the *logical problem*, according to which both mindreaders and behavior-readers have to respond to the *same* stimuli—in this case, features such as body posture, head orientation, facial expression, and the relationship between the bodily posture and environmental features such as the presence of food. The behavior-reader makes an inference from these observable features plus knowledge of behavioral regularities to future behavior. The mindreader must, as well, make an indirect inference to a hidden mental state before judging future behavior, as follows:

(BR): S's behavior at t_1 + knowledge about behavioral regularities → S's behavior at t_2

(MR): S's behavior at t_1 + knowledge about behavioral regularities → S's beliefs and desires → S's behavior at t_2

The problem arises because all of the observable features of the situations are the same in both conditions. To be a mindreader, one must also be a behavior reader.

In both models there is a body of knowledge (sometimes described as a theory) that permits the inference from a belief about another's current behavior to a belief about another's future behavior. Povinelli and Vonk describe this knowledge as follows:

> (1) a database of representations of both specific behaviors and statistical invariants which are abstracted across multiple instances of specific behaviors; (representations that may be formed either by direct experience with the world, or may be epigenetically canalized);
>
> (2) a network of statistical relationships that adhere between and among the specific behaviors and invariants in the database;
>
> (3) an ability to use the statistical regularities to compute the likelihood of the specific future actions of others (Povinelli and Vonk, 2004, p. 6).

The two points I want to keep in mind at this point is that as presented, both Sober's (M) and (B) have intervening variables, and both Povinelli and Vonk's (BR) and (MR) require knowledge/theory about behavioral regularities.

In order to overcome the logical problem and uncover evidence in favour of the existence of the additional common cause—a mindreading representation—Sober suggests that researchers create an two-winged experiment that examines the relationship between two tasks that should be positively associated if there is an intervening mindreading variable that serves as a common cause. The experiment he proposes is a combination of Melis et al. (2006) tunnels and trapdoors tasks. In the tunnel task, chimpanzees were given the opportunity to steal food via an opaque and transparent tunnel, and they overwhelmingly choose to use the opaque tunnel. In the trapdoors task, chimpanzees were given the opportunity to steal food from a noisy and a quiet trapdoor, and they overwhelmingly chose the quiet trapdoor. Melis and colleagues think that each of these experiments offer evidence in favor of a mindreading hypothesis. Logical problem worries arise, however, for the chimpanzee subjects may have formed associations between silence and successful stealing, and lack of direct-line-of-gaze and successful stealing. What Sober suggests, however, is that on the behavior-reading hypothesis we should expect no relationship between these two associations because they do not share the same perceivable property; in order to know that they are connected, the chimpanzee would have to form a unifying thought such as "If I use the noisy door or the transparent tunnel, the competitor will notice me." Mindreading is supposed to be unified, not fragmented, while behavior-reading is supposed to be much less unifying.

By offering chimpanzees the opportunity to steal food quietly and invisibly within the same timeframe, Sobor suggests we can form a prediction about the existence of an additional common causes in a mindreading representation compared with a prediction about a lack of an additional common cause in individuals who lack a mindreading representation. That is, if chimpanzees mindread, then there should be no screening off between tunnel behaviors and trapdoor behaviors, rather there should be a positive correlation between using silent trapdoors and opaque tunnels. However, if chimpanzees do not mindread, there should be screening off between the two tasks—no correlation, conditional on the stimuli. Put informationally, if chimpanzees mindread, the fact that they pass one mindreading task should affect the probability they pass another task—we can make predictions about how they will behave in other mindreading contexts. But if the probability of passing one mindreading task doesn't affect the probability of passing another one, then the responses screen off and provide evidence against the existence of the same intervening variable.

Sober offers a story about the unique causal work of mindreading abilities—or the added value a mindreader has over a behavior reader. He also makes sense of the unification intuition that Tomasello and Call (2006) express when they claim behavior reading requires a different explanation for each task, whereas mindreading permits the same explanation to serve in each case. The essential premise is that mindreading, but not behavior reading, should permit multiple

effects that are not screened off from one another, whereas the opposite should be true if the behavior reading hypothesis is true.

My worry is that both the mindreading hypothesis and the behavior reading hypothesis are, as they are discussed in the literature, catchall hypotheses that are not explicit about their parameters and do no disagree about screening off. Returning to Povinelli and Vonk's (2004) description of the two competing psychological systems at stake, we see that the behavior-reading hypothesis hypothesizes that a theory of behavior is required, and the mindreading hypothesis hypothesizes a theory of behavior plus knowledge of mental states. While Povinelli and Vonk don't specify what a theory of mind would look like, they say that the human theory of mind system "uses information about ongoing, recent, or even quite temporally distant behaviors, to generate inferences about the likely mental states of others' (Povinelli and Vonk, 2004, p. 6). According to Povinelli and Vonk, the attribution of the mental state often merely "go along" with the behavioural regularity, adding little predictive value (Povinelli and Vonk, 2004, p. 9)—though they do think that 'it is possible to imagine situations in which responding appropriately in relatively novel situations might be facilitated by a system that reasons about mental states' (Povinelli and Vonk, 2004, p. 10).

If mindreading doesn't permit much in the way of predictive value, what use is it? Povinelli's Reinterpretation Hypothesis suggests that while humans and apes both predict behaviors in terms of behavioral abstractions, humans also developed a theory of mind system that allowed us to "explain in terms of mental states" (Povinelli and Vonk 2003, 158). An understanding of mental states permits greater understanding of present behavior, but this need not necessitate greater predictive accuracy regarding future behavior.

For example, if a chimpanzee regularly shrinks away from a dominant's display, behaviour-reading chimpanzees can form a regularity between behaviors. They need not know why the chimpanzee engages in this behavior. A mindreader can think about the mental states of the other chimpanzee, and could then suppose that the chimpanzee subordinate shrinks away *because* he is afraid of the dominant.

The worry that the two hypotheses don't disagree about screening off stems from the commitment to the existence of a theory of behavior in the behavior-reading case, and then an added understanding of the mental states that go along with the behaviors in the behavior-reading case. Povinelli and Vonk's treatment of the two psychological models take both to be unified. Both systems require the database of behaviors' statistical relations to other behavior. The mindreading system only adds additional meaning to the behavioral regularities—a mindreader realizes that direct line of gaze is also seeing, and that aural information is heard, and that things that are seen and heard are noticed.

Returning to Sober's proposal, and given the models Povinelli and Vonk propose, we should draw vertical arrows between Tunnels and Trapdoors in the behavior-reading hypothesis part of the diagram on Sober's Fig 4.7. As Sober points out, however, doing so would make the mindreading and behavior-reading models empirically indistinguishable.

Given the opacity of internal mental states and a healthy commitment to the razor of silence, one might not expect a behavior-reader to add vertical arrows

between opacity and silence in Sober's diagram. To defend the claim that they would, consider first why there are vertical arrows on the mindreading hypothesis. What unifies these capacities on a mindreading hypothesis is *an additional intervening variable*. Sober writes, "if a mind-reading chimpanzee has a belief about what the human experimenter can and cannot see in the tunnel task that takes place at a given time, this should raise the probability that the chimpanzee will have a belief about what the human experimenter can and cannot hear in the trapdoor task that occurs immediately thereafter…A mindreader has the resources to apprehend these connections…the one state of believing would raise the probability of the other…a purely behavior-reading chimpanzee will lack the resources for drawing this conclusion" (224-225).

Sober's comment here raises the question of why and how the two beliefs will be connected. What justifies our positing a probabilistic relation between intervening variables are the same behavioral contingencies that permit learning of the mindreading rule, which, by hypothesis, are *exactly the same* as those that permit the learning of the behaviour-reading rule. Thus, if the seeing and hearing rules are related conceptually in the mindreader, it is only because they are related, experientially, in the behavior-reader. Both models require a further relation that unifies the beliefs. The mindreading conceptual unifier might be something like *notice*—both seeing and hearing lead to noticing. If that is right, then the explanation of the arrows is going to be something like what Sober sketches in Fig. 4.8, which is a model that doesn't entail screening off. The reason why the use of the opaque tunnel and the silent trap door are positively correlated is in both cases the actors recognize that the competitor could notice them. And, if this is right, the two-winged task wouldn't be appropriate, as Sober notes.

Similarly, the theory of behavior that Povinelli and Vonk propose would also unify the variables recruited in succeeding in tunnel and trapdoor tasks. If we take their claim in the 2004 paper at face value, it is only in novel situations that behavior rules will have no play. Since observations of behavior already permit generalizations to the same type of situations, there is no reason why these generalizations cannot be unified via a non-mental concept.

A prima facie worry about this proposal is that there may be differences in the behavior reading and mindreading sets of information; mindreading models have fewer adjustable parameters, since behavior reading needs to posit separately learned associations (back/facing; opaque/not opaque; transparent/not transparent). However, given Robert Lurz's development of behavior-reading, these apparently separately learned associations can all be learned in terms of a single concept—direct-line-of-gaze (Lurz 2011). Just as a behavior-reader can group together states of affairs such as putting one's hand in front of one's fear-grin, having a barrier between the dominant and the food, and dodging behind a rock as cases of not being in direct-line-of-gaze (a mindreader might call this not seeing), a behavior-reader could group those states of affairs along with not picking the noisy trap-door and refraining from uttering a food call in one category (a mindreader might call this not noticing).

What would be a non-mental concept that unifies cases of seeing and hearing—doing the work *notice* does for the mind-reading case? Perhaps we can

unify the behavior-reading hypotheses in terms of *detection*. Detection can be glossed as responsiveness to the presence or absence of a certain feature of the world. A piece of litmus paper is a detector, and not minded. A behavior-reader can likewise see that the competitor in the tunnel case would detect in the transparent tunnel and in the noisy trapdoor, given the learning history which of the sorts of stimuli that lead to detection. The behavior-reader thinking is as follows: in the past the times in which I was successful in gaining food was when I wasn't detected, as when I was hidden behind a rock, and when I supressed my typical food call. I should make sure I'm not detected, so I will use the opaque tunnel and the silent trapdoor.

So, while the theoretical information used to predict behavior would be unified for mindreaders who had a learning history in which they were exposed to both noises and visual cues (and had functional sensory systems for both modalities), and the same would be true for the behaviour reader. Akaike model selection criteria would conclude that the mindreading hypothesis would have an advantage over behavior-reading hypothesis only if the later "must posit multiple learned contingencies to explain [model] behavior" (424). But since the learning histories should be the same for group living individuals like chimpanzees, the models are going to posit the same number of learned contingencies.

There are several ways to respond to these worries. One would be to create a different kind of two-winged test that predicted behaviors of a different sort. According to the reinterpretation hypothesis the added value of mindreading is explanation, not enhanced prediction. To determine whether or not chimpanzees mindread, given this difference, would be to examine whether predictions and mental explanations screen off from one another. Povinelli and Vonk explicitly claim that the two hypotheses disagree about this issue. However, we are immediately faced with the worry that chimpanzees can't communicate any possible explanation they might have for behavior, since explanation-giving is canonically understood as verbal behavior. Such a strategy could work with verbal human children, and could certainly assist in adjudicating the current debate about mindreading in children that arises due to the ability of infants to pass implicit false belief tasks, and the inability of young children to fail explicit verbal false belief tasks (see, e.g. Baillargeon et al. 2010).

Another possible response to my worry is to point out that Povinelli and Vonk do think there are some kinds of predictions we should expect a mindreader but not a behavior-reader to make, and that we could make a two-winged experiment that includes one of the tasks that that Povinelli and Vonk think only a mindreader could pass—the goggles task.

The goggles task, which was first proposed by Cecelia Heyes (1998), requires a subject to experience a novel situation first hand, and have no observational knowledge of others' behavior in that situation. If the subject could predict how another individual would act in that situation, they would be engaged in experience projection, and be thinking that their mental states were shared by the target. Specifically, the goggles task involves giving a chimpanzee experience with different invisible properties of two objects that differ only by a color marker, such as two pairs of goggles, one that is transparent and the other opaque. If a chimpanzee

learns from wearing the red goggles that they are transparent, and that the blue goggles are opaque, and if the chimpanzee understands seeing, he could predict that anyone else who wears the red goggles would be able to see as well. And, if the chimpanzee also knows that he can steal from those who can't see, and should beg for food from those who can see, the chimpanzee should take the appropriate actions.

A two-winged version of this task could furthermore involve introducing yellow transparent goggles and green opaque ones, but the chimpanzees learn about the affordances of these goggles in a social setting; they observe that when others put on the green goggles they bang into walls and walk like a zombie, and when they put on the yellow goggles they act normally. (This entire story ignores the fact that chimpanzees would not wear goggles in any of these scenarios, but please permit me some absurdity.) Then, in the test situation the chimpanzee could be given the chance to steal from four individuals, 2 at a time, paring the goggle sets learned under same conditions. If Povinelli and Vonk are right and the mindreader would reliably choose both the red and the yellow goggles but a behavior-reader would reliably choose only the yellow goggles, then we could analyze the cases in terms of screening off as Sober recommends.

There is one big problem with this suggestion, though, namely that Povinelli and Vonk are probably wrong that only a mindreader would reliably choose the red goggles. As Sober notes in a footnote, I argued that the goggles experiment cannot decide between the two hypotheses as they have been presented both here and in Sober's chapter (Andrews 2005), and none of the subsequent attempts to design hypotheses have improved the situation (Andrews 2012). This is because experience doesn't entail knowledge of mental state, but experience can provide information about behavioral affordances. The chimpanzees who put on the opaque blue goggles might not realize they cannot see, but realize that they cannot do things, and hence choose to steal from someone who later wears blue goggles, and avoid from begging from someone who is wearing blue goggles, because blue goggle wearers can't do things.

A final response to this worry about the hypotheses challenges my suggestion that the behavior reader's social models are just as unified as the mindreader's models in virtue of there being nonmentalistic intervening variables rather than mentalistic intervening variables. Perhaps, some might think, the mere existence of unifying intervening variables would be sufficient to overcome the logical problem, and on my reading the logical problem is unreasonably strong. According to this strong interpretation, there needs to be evidence that the intervening variable has a specifically mentalistic sense—not just referent (e.g., Lurz 2011). A weaker interpretation assumes we only need show that the chimpanzee uses some intervening variable between stimulus and response (e.g. Whiten 1996). In that case, evidence of either the intervening variable of detecting or the intervening variable of knowing both provide evidence for the mindreading hypothesis.

A weaker version of the logical problem also arises in Penn and Povinelli (2007), where, in responding to criticisms of the previous position, they admit that evidence of any intervening variable at work would be sufficient evidence for

mindreading; they write, "being able to recode perceptually disparate behavioral patterns resulting from the same underlying cognitive state as instances of the same abstract equivalence class is a bona fide example of postulating an ms variable in the sense defined hereinabove" (Penn and Povinelli 2007, 733). In response to my alternative explanation of what might be going on in the goggles case, they write:

> In order to infer that the experimenter is not likely to respond to begging gestures while wearing the red visor, the subject must realize that responding to begging gestures requires more than a set of manifest physical actions and observable conditions. To be precise, the subject must realize (by logical inference or embodied simulation, or some combination of the two) the following:
>
> (i) wearing the opaque visor results in an inability to 'see-what-is-going-on' (i.e. a general epistemic condition applicable to any subsequent behavior not just a particular manifest physical effect of bumping-into-things),
>
> (ii) this general epistemic condition will be experienced, analogously, by the other subject when she wears the red visor but not the blue visor, and
>
> (iii) a subject who experiences this general epistemic condition will not respond to begging gestures. (Penn and Povinelli 2007, 738).

What they call an epistemic condition could just as well be stated as an informational condition, as I do for Sober's proposed two-winged tunnels and trapdoors case. In fact, what the content of the intervening variable amounts to is less important than the mere case that there is an intervening variable that organizes stimuli-response patterns functionally. (Indeed, a functionalist about mental states would be hard-pressed to deny that the ability to track the relations between what we see as mental states and action is sufficient as evidence of mental state attribution.)

If mere evidence of an intervening variable is sufficient for evidence for mindreading, then not only do we not need a two-winged task, but the goggles test could provide the needed evidence. And, recently, it has. Chimpanzees (Karg et al. 2015) and ravens (Bugnyar et al. 2016) have have passed versions of the goggles task.

However, a weak reading of the logical problem creates additional problems for testing the hypotheses. If we return to how Sober states the two hypotheses:

> (M) D's behavior → S's beliefs about D's behavior → S's beliefs about D's mind → S's behavior
>
> (B) D's behavior → S's beliefs about D's behavior → S's behavior

we recall that there is an intervening variable in *both* the mindreading and behavior

reading hypothesis, because no one in the debate denies that chimpanzees are minded. As Penn and Povinelli put it, the starting point is "that cognitive agents—biological or otherwise—can learn from their past experience, in part because they have dynamic internal states that are decoupled from any immediate physical connection to the external world" (Penn and Povinelli 2007, 732). If the hypotheses are presented in terms of intervening variables already, then the only evidence for (M) over (B) would turn on S's beliefs about D's mind providing some unique causal work over and above beliefs about behavior. The theory of behavior that Povinelli and colleagues build into their model is so rich and robust that it is going to be behaviorally indiscernible from a mindreading model. Even language won't help, assuming that linguistic behavior could be gained via experience in a linguistic community by a nonmentalistic system. The logical problem artificially separates one's knowledge of behavior and one's knowledge of mind. If we reject this form of dualism, the problem doesn't arise. The knowledge of behavior regularities just is knowledge of mental states in certain cases, and the concept that is used to categorize the behavior types together just is a mental state concept. The behavior-reading hypothesis doesn't snip away an intervening variable, rather the mindreading hypothesis edits the content of the knowledge store of behavioral regularities.

[The problem doesn't come from Sober's statement of the hypotheses, but in the hypotheses and theoretical structure built around them. The behavior-reading hypothesis is a bit of shifting sand that doesn't manage to stay still for very long. I was sitting next to Vonk when Sober presented an early version of this argument at the Southern Society for Philosophy and Psychology. "We don't think that," she whispered to me during the talk. The problem is, it isn't exactly clear what they do think.]

Andrews, K. (2012). Review of *Mindreading Animals: The Debate over What Animals Know about Other Minds*. Retrieved from http://ndpr.nd.edu/news/29824-mindreading-animals-the-debate-over-what-animals-know-about-other-minds/

Andrews, K. (2005). Chimpanzee theory of mind: Looking in all the wrong places? *Mind and Language*, *20*, 521–536.

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants.*Trends in Cognitive Science*,*14*, 110–118.

Bugnyar, T., Reber, S. A., & Buckner, C. (2016). Ravens attribute visual access to unseen competitors. *Nature Communications*, *7*, 10506.

Heyes, C. (1998). Theory of mind in nonhuman primates. *Behavioral and Brian Sciences*, *21*, 101–134.

Karg, K., Schmelz, M., Call, J., & Tomasello, M. (2015). The goggles experiment: can chimpanzees use self-experience to infer what a competitor can see? *Animal Behaviour*, *105*, 211–221.

Lurz, R. (2011). *Mindreading Animals*. Cambridge, MA: MIT Press.

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

Povinelli, D. J., & Vonk, J. (2004). We don't need a microscope to explore the chimpanzee's mind. *Mind and Language*, *19*, 1–28.
Whiten, A. (1996). 17 When Does Smart Behaviour-Reading Become Mind-Reading? In P. Carruthers & P. K. Smith (Eds.), *Theories of Theories of Mind* (p. 277). Cambridge University Press.