Inference and Error in Comparative Psychology: The Case of Mindreading[1]

Marta Halina

10 July 2015

Abstract

Mindreading is the ability to attribute mental states to other agents. Over the last decade, there has been a wealth of experimental work on the question of whether nonhuman animals mindread. The positive results of these experiments have led many comparative psychologists to conclude that animals attribute some mental states, such as intentions and perceptions, to others. Sceptics remain, however. They argue that one can provide alternative non-mindreading explanations for the positive results of mindreading experiments, and that insofar as this can be done, the hypothesis that animals mindread lacks evidential support. In this paper, I argue that this "alternative-hypothesis objection" depends on an oversimplified view of the relationship between theory and evidence. A more nuanced account reveals that the mindreading hypothesis is supported by the data produced by mindreading experiments, while alternative hypotheses, such as behaviour reading and submentalizing, lack such support. I conclude by considering whether these alternative hypotheses undermine the evidence for mindreading by serving as experimental confounds. I argue that their ability to do so depends on their independent evidential support and that mindreading sceptics have not done enough to show that they have such support.

1. Introduction

Mindreading is the ability to attribute mental states to other agents. It is what we do when we predict and explain the behaviour of others by appealing to their beliefs, desires, intentions, and perceptions, rather than just their observable behaviour. Mindreading is thought to be ubiquitous in adult human life and to underlie many other cognitive abilities, such as empathy, self-awareness, and even phenomenal consciousness (Baron-Cohen 1997; Carruthers 2009; Apperly 2011). Many of these abilities have long been held to be uniquely human. Discovering whether nonhuman animals mindread then would dramatically affect not only how we view them, but also how we view ourselves.

Psychologists and philosophers have been pursuing the question of whether nonhuman animals (hereafter, animals) mindread for over 35 years (Premack and Woodruff 1978). From the beginning of this research program, there has been a debate over how to interpret the positive results of mindreading experiments. On the one hand are those psychologists and philosophers who take these positive results as good evidence for animal mindreading (Call and Tomasello 2008; Fletcher and Carruthers 2013; Halina 2015; Clayton 2015); on the other are those who do not (Povinelli and Vonk 2006; Penn et al. 2008; Penn and Povinelli 2007, 2009, 2013; Penn

[1] A previous version of this paper was presented to the Cambridge Comparative Cognition Lab. Thanks to the members of that group for their helpful feedback and discussion, especially Lucy Cheke, Nicky Clayton, Ed Legg, Corina Logan, and Ljerka Ostojic. A very special thanks to Kristin Andrews, Irina Mikhalevich, and Robert Lurz for agreeing to comment on this paper for Minds Online—thank you!

2011; Lurz 2011; Heyes 2014, 2015; Buckner 2013). I refer to the latter group as *methodological sceptics* (or "sceptics") given their doubt in the adequacy of the methods currently used to test for mindreading in animals.[2]

The "alternative-hypothesis objection" is central to the critique advanced by the sceptics.[3] This objection holds that if an alternative non-mindreading hypothesis can account for the results of a mindreading experiment, then those results do not in fact provide good evidence for mindreading. In such cases, both the mindreading and alternative hypotheses are equally well supported by the data. The sceptics often go on to conclude that we should accept one of the alternative hypotheses on the grounds of it being simpler than positing mindreading. The focus of this paper is on the first two claims, however.[4] In particular, I address the question, "does the ability to account for the results of a mindreading experiment with an alternative hypothesis (1) undermine the evidential support for the hypothesis being tested and/or (2) provide good evidence for the alternative?" My answer to both parts of this question is "no."

Philosophers of science have been thinking about the relationship between theory and evidence for a long time, and in particular, when data should count as good evidence for a hypothesis. Few of those engaged in the mindreading debate, however, have drawn on general philosophy of science in order to better understand the alternative-hypothesis objection. I do this here by evaluating this objection within the framework of experimental testing and inference developed by error-statistical philosophers like Deborah Mayo (Mayo 1996, Mayo & Spanos 2008, Staley 2008). I argue that the alternative-hypothesis objection depends on an over-simplified view of the relationship between theory and data. This view holds that it is sufficient for data to "fit" or be consistent with a hypothesis in order for it to serve as good evidence for that hypothesis. I instead argue that in order for data to serve as good evidence for a hypothesis, it must be produced by a "severe test" or a testing procedure capable of detecting whether that hypothesis is false when it is in fact false. With this additional criterion for good evidence in place, I show that mindreading experiments constitute severe tests with respect to the mindreading hypothesis being tested, but not with respect to the alternative hypotheses of behaviour reading (Penn and Povinelli 2007) and submentalizing (Heyes 2014, 2015). Thus, although the positive results of mindreading experiments are consistent with both mindreading and these alternatives, they provide good evidence only for the former.

I begin in section 2 by introducing the logic behind mindreading experiments and why comparative psychologists take them to provide good evidence for mindreading. In section 3, I introduce the alternative-hypothesis objection. I then show in section 4 how this objection depends on an overly permissive account of good evidence and introduce severe testing as a corrective for this view. In section 5, I argue that mindreading experiments are severe tests with respect to the mindreading hypothesis, but not with respect to the alternatives of behaviour reading and submentalizing. I conclude in section 6 by considering one way in which the alternative-hypothesis objection might succeed in weakening the evidence for mindreading: by flagging experimental confounds. In order to do this, however, sceptics must show that these

[2] A somewhat similar distinction has been made between "romantics" and "killjoys" and "boosters" and "scoffers" (Dennett 1983; Tomasello and Call 2006; Shetttleworth 2010).
[3] I borrow this term from Mayo 1996 (see below).
[4] See Sober 2001, Lurz 2011 (especially section 2.7), and Meketa 2014 for discussion of the latter claim.

purported confounds have independent empirical support—something that they do not currently do.

2. Mindreading Research

How do comparative psychologists test whether animals attribute mental states to other agents? Although the particular methods vary depending on the mental state in question, the general approach used in mindreading experiments is the same.[5] Indeed, the logic follows that of experimental design in psychology more generally.[6] In order to see this, it is useful to first distinguish between two kinds of hypotheses: high-level and experimental.[7] In psychology, the *high-level hypothesis* is typically the cognitive account or mechanism that researchers are testing through the implementation of a battery of experiments and observational studies. These are the claims that chimpanzees understand intentions or have level 1 visual perspective taking abilities. Such high-level hypotheses give rise to a range of concrete predictions. It is these predictions that serve as the basis for experimental hypotheses, where an *experimental hypothesis* is a claim that some factor will vary with another factor (what will become the dependent and independent variables in an experiment). Such hypotheses include the claim that subjects will prefer to steal food from a competitor by reaching through an opaque tunnel over a transparent one (Melis et al. 2006) or request food from a cooperative agent by gesturing towards their front rather than their back (Liebal et al. 2004).

Experimental hypotheses are then tested by means of a well-designed experiment or one that is internally valid and set up to test the relationship between the two variables in question. The independent variable is the variable manipulated by researchers across conditions (such as the opaqueness of a barrier), whereas the dependent variable is that which researchers predict will be affected by the independent variable in a particular way (such as a subject's attempt to retrieve food). Ensuring internal validity requires that researchers control nuisance variables or those factors that might affect the dependent variable other than the independent variable. Nuisance variables need to be eliminated or randomized in order to prevent them from systematically affecting the dependent variable. A nuisance variable that has such a systematic effect may either produce a difference across conditions that researchers mistakenly attribute to the independent variable or produce an effect counter to the independent variable, thereby masking the latter's impact. When nuisance variables are randomized, their effects are taken into account in the statistical analysis of the data. If the results of the experiment are statistically significant and match the prediction, then this is taken as positive evidence for the experimental hypothesis under test.

When a large number of experimental hypotheses are confirmed, this is taken as evidence in favour of the high-level hypothesis predicting them.[8] Generally, the greater the number and variety of confirmed experimental hypotheses, the more confident researchers are in the truth of

---

[5] See Premack and Woodruff 1978, Povinell et al. 1990, and Povinelli and Eddy 1996 for pioneering work in this area.

[6] See Sani and Todman 2006 and Dienes 2008 for general introductions to experimental design.

[7] See Mayo 1996 and Staley 2008. Sani and Todman 2006 also refer to the former as a "theory" and the latter as a "testable hypothesis."

[8] I use "confirm" loosely here to mean, "judged to be supported by the data."

the high-level hypothesis. It is for these reasons that comparative psychologists, such as Nicola Clayton, Michael Tomasello, and Josep Call hold that animals are capable of some forms of mindreading. The predictions made by the high-level hypothesis that chimpanzees are capable of level 1 visual perspective taking, for example, has led to many positive results. These results support the high-level hypothesis, according to proponents of the current experimental approach. Let us now turn to the objection posed by the sceptics.

3. The Alternative-hypothesis Objection

The methodological sceptics argue that the above experiments do not in fact provide good evidence for mindreading because there are alternative non-mindreading hypotheses that can account for the experimental results. Over the last decade, Povinelli and colleagues have argued that these results can be explained by a "behaviour-reading hypothesis" (Povinelli and Vonk 2006; Penn et al. 2008; Penn and Povinelli 2007, 2009, 2013; Penn 2011; see also Lurz 2011 and Buckner 2013); while, recently, Cecilia Heyes has argued that they can be accommodated by a "submentailizing hypothesis" (Heyes 2014, 2015). In both cases, the sceptics hold that insofar as the results of mindreading experiments can be explained by these alternative accounts, they do not provide good evidence for mindreading. Instead, because the results are consistent with both mindreading and these alternatives, and the alternatives can be viewed as simpler than mindreading, it is these alternatives that researchers should accept.

Penn and Povinelli (2007) represent a clear example of this argument. They write that, "in order to produce experimental evidence for an f_{ToM} [theory of mind function], one must first falsify the null hypothesis that the agents in question are simply using their normal, first-person cognitive state variables" (734). They then show how one can construct an alternative non-mindreading explanation for every positive result produced by mindreading experiments. Constructing such an explanation involves more or less positing that subjects have a rule that links whatever observable variable the experimenter is manipulating with whatever dependent variable is being measured. The claim that subjects rely on such a collection of rules is what constitutes the behaviour-reading hypothesis. Given that the results of mindreading experiments are consistent with this alternative, Penn and Povinelli conclude, "the available evidence suggests that chimpanzees, corvids and all other non-human animals only form representations and reason about observable features, relations and states of affairs from their own cognitive perspective" (737).

Heyes's general strategy is similar to Povinelli and colleagues. She writes, "all of the results published in recent years are subject to the observables problem; they could be due to mindreading, but they are at least equally likely to reflect exclusive use for social decision-making of directly observable features of the stimulus context" (Heyes 2015, 316). Heyes departs from Povinelli and colleagues, however, in advancing submentalizing as the alternative hypothesis of choice. According to this hypothesis, subjects solve mindreading tasks by employing "domain-general cognitive processes that do not involve thinking about mental states but can produce in social contexts behavior that looks as if it is controlled by thinking about mental states" (Heyes 2014, 132). Insofar as domain-general processes such as memory, attention, and perception can account for the positive results of mindreading experiments, Heyes

argues, they do not provide evidence for mindreading. Heyes then shows how the best mindreading experiments can be reinterpreted in this way.

4. An Additional Criterion for Good Evidence: Severe Tests

The problem of being able to accommodate a set of data by multiple theories or hypotheses is well known in philosophy of science. Deborah Mayo refers to this as the "alternative-hypothesis objection" or "methodological underdeterminism." She characterizes the general objection as follows: "Evidence in accordance with hypothesis *H* cannot really count in favour of *H*, it is objected, if it counts equally well for any number of (perhaps infinitely many) other hypotheses" (1996, 174).

Many philosophers now agree, however, that this objection rests on an oversimplified view of the relationship between theory and evidence. Namely, that in order for data to serve as evidence for a hypothesis, it need only be consistent with that hypothesis. The problem with this view can be illustrated with a simple example. Imagine that you would like to test the claim (H_1) that running a marathon will not lead some person *x* to lose more than five kilograms of weight. To test this, you weigh *x* before and after the marathon, finding that the scale indicates 50 kilograms both times. The data fits your hypothesis and seems like good evidence for it. Now imagine that you want to test the claim (H_2) that running a marathon will not lead *x* to lose more than half a kilogram of weight. You conduct the same test and find again that the scale reads 50 kilograms both times. This data also fits your hypothesis and seems like good evidence for it. That is, until you discover that the scale that you have been weighing *x* on is sensitive to changes of weight of only one kilogram or more. The collected data still fits H_1 and H_2, but most would contend that you only have evidence for H_1. This is because if H_2 were false, if *x* lost more than half a kilogram of weight after the marathon (600, 700, 800, 835 grams, for example) this particular scale would be unable to detect it. In contrast, if H_1 were false, if *x*'s weight had changed by five kilograms or more, your measuring instrument would have indicated this by producing data that is discordant with H_1.

This example captures the idea that producing data that fits one's hypothesis is not enough for it to serve as good evidence for that hypothesis. Instead one must take the test procedure used to produce that data into account and ensure that it is capable of producing data that is discordant with that hypothesis when that hypothesis is in fact false. Such test procedures are what Mayo (1996) calls "severe" or "error probing." Recognizing this additional constraint on good evidence dulls the threat of methodological underdetermination. The ability to conceive of alternative hypotheses consistent with one's data is not enough to show that these alternatives are serious rivals with evidential support. One must also show that the tests responsible for producing that data are severe with respect to those alternatives.[9]

[9] One can formally evaluate the severity of a test by calculating the probability of obtaining a given set of data on the assumption that the hypothesis being tested is false. However, I will undertake a more qualitative analysis here. A formal analysis is not always possible with high-level hypotheses and the particular statistical methods used to evaluate experimental hypotheses in psychology often vary from experiment to experiment. Given this, a general, qualitative analysis is more appropriate and, as we will see, adequate for evaluating the claims being considered here.

5. Mindreading and its Alternatives: Are They Severely Tested?

Severe tests provide an additional constraint for evaluating evidence for a hypothesis. In this section I apply this general lesson to the alternative-hypothesis objection advanced by mindreading sceptics. I argue that mindreading experiments are severe with respect to the mindreading hypothesis being tested, but not with respect to the alternative behaviour-reading and submentalizing accounts. If the behaviour-reading and submentalizing hypotheses are to be considered genuine rivals to mindreading, we must look elsewhere for their empirical support.

5.1. Mindreading Hypotheses

As we saw in section 2, a mindreading hypothesis is a high-level hypothesis that makes predictions, which serve as the basis for experimental hypotheses. In order to determine whether a high-level mindreading hypothesis has been severely tested, we must look at both stages of this process.

Let us begin with experimental hypotheses. An experimental hypothesis has been severely tested when an experiment is internally valid and the results are statistically significant. Recall that an experimental hypothesis in a mindreading experiment takes the form of a claim that there will be a difference in the dependent variable between two conditions—one in which the independent variable is present and one in which it is absent. Data that is discordant with this hypothesis would take the form of no observed difference between conditions. Now in order to determine the severity of this test, we must ask, how likely is it that it would produce such discordant data if the experimental hypothesis were false? If researchers follow standard protocol for experimental design and statistical analysis, then this likelihood is high.

As we saw, standard experimental protocol dictates that researchers control for nuisance variables by holding them constant or randomly allocating them across conditions. Successfully doing this means that the only systematic change in the dependent variable—if there is one at all—will likely be due to the independent variable. Randomized nuisance variables may accidently produce systematic effects by piling up in one condition or another, but this possibility is taken into account in the statistical analysis. Within-condition variance of the dependent variable is used to gauge how much noise is being produced by randomized nuisance variables.[10] This variance is used to determine the threshold or significance level at which one should accept the experimental hypothesis. The more variance there is, the bigger the difference required between conditions in order to conclude that the effect is due to the independent variable, rather than to the randomized nuisance variables.

In controlling for nuisance variables in this way, researchers are maximizing the probability that the data will be discordant with the experimental hypothesis, if that hypothesis were false. They are creating a situation in which, if the independent variable had no effect on the dependent variable, it would be unlikely that the data leading to the acceptance of this claim (i.e., a statistically significant difference) would be produced. Formally, for a given statistical test, this probability is 1-α, where α is the probability of falsely accepting the experimental hypothesis

[10] Recall that the independent variable does not vary within a condition, so all within-condition variance should be due to randomized nuisance variables.

(that is, accepting that the independent variable produced the observed effect when in fact it did not). In psychology experiments, α is typically set to less than 5%, which means that the severity of these tests or the probability of rejecting the experimental hypothesis when it is false is greater than 95%. In other words, such tests are almost maximally severe.

The above concerns experimental hypotheses, but how does this affect the evaluation of a high-level mindreading hypothesis? The high-level hypothesis is the source of the claim that a relationship will be found between the independent and dependent variables being tested. By making such a prediction, and subjecting it to testing, the high-level hypothesis is putting itself at risk. Each experiment is an opportunity for discordant data to be produced in the form of a negative result. The more predictions that a high-level hypothesis makes and tests, the more likely it is that discordant data will be produced, if that hypothesis is false. In this way, high-level mindreading hypotheses, such as the claim that chimpanzees have level 1 visual perspective taking abilities, have been severely tested. The chance of all of those results aligning themselves in precisely the way that the mindreading hypothesis predicts, while the mindreading hypothesis is false, is unlikely.[11]

Of course, although the production of all of this concordant data is unlikely if the mindreading hypothesis were false, it is still possible. Perhaps the mindreading hypothesis accurately predicts what other animals will do in a variety of social situations, even though the hypothesis itself is wrong. Perhaps the world is organized in such a way that subjects behave as we expect mindreading agents to behave, but for reasons that have nothing to do with mindreading. These are possibilities, but they are unlikely. The number of predictions made by the mindreading hypothesis, the number of experiments run to test these predictions, how carefully controlled these experiments are—all of these things add to the severity by which a given mindreading hypothesis has been tested. When they are present, the evidence for that hypothesis is good.

5.2. The Alternative Hypotheses

Let us now turn to the alternative hypotheses of behaviour reading and submentalizing. These hypotheses fit the data produced by mindreading experiments, but are the data good evidence for them? Starting with experimental hypotheses, one might be tempted to run the same analysis as above, but an immediate problem arises. The severity by which a particular experimental hypothesis is tested matters in the mindreading case because a negative result counts as discordant data with respect to the high-level mindreading hypothesis. However, this is not the case for the alternative hypotheses of behaviour reading and submentalizing. For these alternatives, both the positive and negative results of mindreading experiments are consistent with their accounts. Indeed, proponents of these alternatives count such negative results as evidence for their views (Penn & Povinelli 2007, Heyes 2015).

That both the positive and negative results of mindreading experiments are consistent with behaviour reading and submentalizing indicates that these experiments are not severe tests with respect to these hypotheses. The fact that these experiments are internally valid and produce

[11] A worry here is that positive experimental results in psychology (and science generally) are over-represented because of the unwillingness of journals to publish negative results. I will set this worry aside here, but it is a legitimate one—thanks to Lucy Cheke for flagging it.

statistically significant results is irrelevant from the perspective of these alternatives because a negative result would not count against them. If these alternatives were false, mindreading experiments would not be able to detect it.

The fact that mindreading experiments are not severe tests with respect to behaviour reading and submentalizing is not surprising when one recognizes that these experiments were not designed to test these hypotheses. The submentalizing hypothesis, for instance, makes claims about how certain environmental objects affect an organism's perception and memory. Heyes (2015), for example, reinterprets the positive results of a visual perspective taking experiment on ravens (Bugnyar 2011) by positing that the memory of subjects was cued in one condition by the presence of a competitor, but not cued in a second condition because the competitor appeared in a different context. This is an interesting claim, but requires an alternative experiment in order to test it—one designed to do so. Such an experiment would require choosing the right independent and dependent variables (ensuring construct validity) and controlling for nuisance variables (ensuring internal validity). For example, given the evidence in favour of visual perspective taking in corvids, one would certainly not want the visual access of a competitor to a food-hiding event to vary across conditions, as it does in this study. And the dependent variable should be a clear indicator of the disruption of "what" and "where" short-term memory. Perhaps a change detection task would be more appropriate than the choice of re-caching one of several food items in the presence of a competitor (Leising et al. 2013). I take it to be widely accepted among experimental psychologists that a well-designed experiment requires stating the hypothesis in advance and designing an experiment to test it rather than some other hypothesis. The claim that the data produced by mindreading experiments do not serve as evidence for behaviour reading or submentalizing is an extension of this point. To produce evidence for these hypotheses, one must conduct experiments designed to do so.

If mindreading experiments were not designed to test behaviour reading and submentalizing, why do they produce data that fits these hypotheses so well? Does this fit not serve as some indicator that these hypotheses reflect the way the world actually is? Here the concept of severe testing is again helpful for diagnosing the situation. Recall that a strength of the high-level mindreading hypothesis is that its predictions have survived severe tests. In contrast, the behaviour-reading and submentalizing hypotheses accommodate data as they are produced (Fletcher & Carruthers 2013). Hypotheses that are constructed on the basis of known data are referred to as "use-constructed" or "rigged" (Mayo 1996). Not all such rigged hypotheses are problematic, but they are problematic when their method of construction minimizes the chances of their being identified as false when they are in fact false. Consider, as an example of this, Mayo's Texas sharpshooter:

> Having shot several holes into a board, the shooter then draws a target on the board so that he has scored several bull's-eyes. The hypothesis, *H*, that he is a good shot, fits the data, but this procedure would very probably yield so good a fit, even if *H* is false (1996, 201).

Such a hypothesis is not only use-constructed, but constructed in such a way that it could not have failed to fit the data, even if it were false. If the behaviour-reading and submentalizing hypotheses are rigged in this way, then their fit with the data says little to nothing about whether

they are likely to be true. Thus, we must ask, if these hypotheses were false, what kind of data would indicate this and do we have the means for producing it? Both the behaviour-reading and submentalizing hypotheses do not fare well in this regard. The reason is that they are both so vaguely specified and flexible that it is not clear whether it is possible to produce data that is inconsistent with them (Fletcher & Carruthers 2013, Halina 2015). For those who hold that it is possible to produce such data (Penn & Povinelli 2007, Heyes 2015), they have not yet constructed the means for doing so. Given this, the current fit that these hypotheses have with the available data is best attributed not to their predictive and empirical success, but to the fact that they are so underconstrained that they can accommodate all of the data that comes their way. Proponents of these views have not done what is required in order to meaningfully test them—that is, test them in ways that are capable of detecting whether they are false. In this case, I would state with John Worrall that, "the 'success' of the theory clearly tells us nothing about the theory's fit with Nature, but only about its adaptability *and* the ingenuity of its proponents" (1989, 155, emphasis original; quoted in Mayo 1996).

6. Alternative Explanations as Experimental Confounds

Mindreading experiments do not provide evidence for the alternative hypotheses of behaviour-reading and submentalizing. Why then do sceptics take these alternatives as undermining the evidence in support of particular mindreading hypotheses? Another possible reason for this is that sceptics take them as threatening the internal validity of mindreading experiments in the form of experimental confounds. Typically, experimental confounds take the form of nuisance variables that are eliminated or randomized, as discussed above. However, sceptics tend to hold that it is precisely the independent variable that is confounding the results of a given mindreading experiment. That is, they often do not dispute that the experiment has established a relationship between the independent and dependent variables, but that the mindreading hypothesis is not the best explanation for this relationship because we have good reason to think that this relationship would hold for reasons unrelated to mindreading. This is a legitimate strategy. If we have good reason to expect a relationship between two variables to obtain independently of the mindreading hypothesis being true, then testing for this relationship is not a good way of testing for mindreading. The question then is whether the behaviour-reading or submentalizing hypotheses constitute good reasons for thinking that the relationships discovered in mindreading experiments would have occurred independently of mindreading.

This is a difficult question because it hinges on what we mean by "good reason." As we saw above, the behaviour-reading and submentalizing hypotheses cannot draw on the results of mindreading experiments for empirical support. The evidence for their plausibility has to come from elsewhere. But exactly how much support or plausibility is needed in order for something to be considered a legitimate experimental confound has not been discussed widely in the literature. Good experimentalists try to control for all factors that might affect the dependent variable, regardless of whether concrete evidence for a particular factor having such an effect has been provided. The psychologists that have been conducting mindreading experiments are no different in this respect. However, these psychologists are right to become wary when a research program is criticised through the positing of a wide range of experimental confounds that were not considered plausible until an experiment produced positive results. When this happens, one must ask what the difference is between constructively flagging experimental confounds and

presenting a sceptical foil aimed at undermining evidence for a hypothesis for the sole sake of undermining it. The latter strategy is undesirable, not least because it violates what Staley (2008) identifies as Pierce's rule: to not "block the way of inquiry" (403). Staley writes that, "to use the mere possibility of error, in the absence of any real doubt, as an obstacle to accepting the result of a sound probable inference, would be to violate Peirce's rule" (403).

How then do we identify a legitimate confound? I propose that in order for a purported confound to be considered legitimate, it should at least have some independent empirical support—and the more support that it has, the more serious it should be taken. In psychology, it is simply too easy to come up with alternative cognitive mechanism as purported confounds. If we allowed every such alternative to be taken seriously by experimentalists, regardless of their support, research would be unable to proceed.

If we understand the behaviour-reading and submentalizing hypotheses as collections of proposed experimental confounds, then these confounds currently vary in their empirical plausibility. With respect to the behaviour-reading hypothesis, early proposals had empirical support. For example, the critique of Hare et al. (2000) that subordinate subjects might avoid the food that the dominant competitor saw because the competitor was allowed to approach that food and so might have simply scared the subordinate away from it was empirically plausible (D'Arcy & Povinelli 2002). Behavioural responses of subordinate chimpanzees to dominants were well known at the time (cite best observational studies). However, over the last ten years, as more positive results on the visual perspective taking abilities of chimpanzees and other animals have come in, and researchers have improved their experimental designs, the purported confounds cited by sceptics have become less plausible. If the confounds cited are no more than an abstract set of innate behavioural rules that have no independent empirical support, then comparative psychologists should not take them seriously.

Heyes's submentalizing hypothesis is advanced as an improvement over the behavioural-reading hypothesis in precisely this respect (Heyes 2015). She argues that a significant weakness of the behaviour-reading account is its lack of empirical support, writing, "the vast majority of behaviour rules considered in current research on mindreading are based on common sense categories… and are not supported or constrained by empirical evidence of any sort" (321). She proposes the submentalizing hypothesis as "a better conception of 'not mindreading'" that is "less dependent on common sense than the current conception of behaviour reading" (322). This moves us in the right direction of focusing on only those experimental confounds that have empirical support. However, the crucial point here is not a move away from "common sense", but a move towards empirically informed alternatives. There is nothing inherently wrong with a hypothesis originating from common sense, as long as that hypothesis has been tested. There are many examples of successful hypotheses with non-scientific origins, such as Kekulé's famous dream-inspired discovery of the structure of benzene. The problem with the behaviour-reading hypothesis is instead that it is currently a collection of untested conjectures.

This aside, the submentalizing hypothesis fares better than behaviour reading because it draws on the known domain-general cognitive abilities of organisms, such as memory, perception, and attention. Even here, though, Heyes does little to show that her proposed confounds are empirically plausible, rather than simply conceivable. For example, she argues that it is

"possible" that the introduction of an opaque barrier prevents a competitor's presence from cuing retrieval from memory the location of hidden food in Bugynar's visual perspective taking experiment on ravens, but cites no studies showing that the introduction of such objects typically has this effect on this subject group. Before requiring that comparative psychologists undertake the laborious and expensive task of rerunning experiments with additional control conditions, sceptics should make a good case for the legitimacy of their purported confounds.

To summarize, the burden should not fall on those psychologists conducting mindreading experiments to show that they can eliminate all non-mindreading hypotheses consistent with their data. This is not possible and would needlessly block research. The burden instead falls on sceptics to show that their purported confounds are not merely sceptical foils. To do so, they must show that these confounds legitimately threaten the internal validity of a particular mindreading experiment by providing independent empirical evidence for their likely presence in the experimental context in question. Pointing to the fact that these alternative explanations fit the data produced by mindreading experiments does not constitute such evidence.

7. Conclusion

Proponents of behaviour reading and submentalizing tend to characterize these accounts as hypotheses that are equally well supported by the data produced by mindreading experiments. This rests on a misconception of what constitutes evidential support for a hypothesis: fit is not enough. The tests producing that data should also be severe or capable of detecting that the hypothesis is false when it is in fact false. Mindreading experiments are severe with respect to the high-level mindreading hypothesis being tested, but not with respect to behaviour reading and submentalizing. In order to provide experimental evidence for the latter, one must test them with experiments that were designed to do so. Such independent evidence is also required if these alternatives are to be taken as legitimate confounds in mindreading experiments.

References

Andrews, Kristin (2012). Do apes read minds? Toward a new folk psychology. Cambridge, MA: MIT Press.

Apperly, Ian (2011). Mindreaders: The cognitive basis of "theory of mind." New York, NY: Psychology Press.

Baron-Cohen, Simon (1997). Mindblindness: An essay on autism and theory of mind. Cambridge, MA: MIT Press.

Buckner, Cameron (2013). The semantic problem (s) with research on animal mindreading. Mind & Language, 29(5): 566-589.

Bugnyar, Thomas (2011). Knower–guesser differentiation in ravens: Others' viewpoints matter. Proceedings of the Royal Society B: Biological Sciences, 278(1705): 634-640.

Call, Josep, and Michael Tomasello (2008). Does the Chimpanzee Have a Theory of Mind? 30 Years Later. TRENDS in Cognitive Sciences 12 (5): 187–92.

Carruthers, Peter (2009). How we know our own minds: the relationship between mindreading and metacognition." Behavioral and Brain Sciences, 32(2):121-138.

Clayton, Nicola (2015). Ways of thinking: From crows to children and back again. The Quarterly Journal of Experimental Psychology, 68(2): 209-241.

D'Arcy, Karen R. M. & Povinelli, D. J. (2002). Do chimpanzees know what each other see? A closer look. International Journal of Comparative Psychology, 15(1): 21-54.

Dennett, Daniel C. (1983). Intentional systems in cognitive ethology: The "Panglossian paradigm" defended. The Behavioral and Brain Sciences, 6: 343-390.

Dienes, Zoltán (2008). Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference. New York: Palgrave Macmillan.

Fletcher, Logan, and Peter Carruthers (2013). Behavior-Reading versus Mentalizing in Animals. In Agency and Joint Attention, ed. Janet Metcalfe and Herbert S. Terrace, 82–99. Oxford: Oxford University Press.

Halina, Marta (2015). There is no special problem of mindreading in nonhuman animals. Philosophy of Science, 82: 473-490.

Hare, Brian, Josep Call, Bryan Agnetta, and Michael Tomasello (2000). Chimpanzees know what conspecifics do and do not see. Animal Behaviour, 59: 771-785.

Heyes, Cecilia. M. (2014). Submentalizing: I'm not really reading your mind. Perspectives on Psychological Science, 9: 131-143.

Heyes, Cecilia. M. (2015) Animal mindreading: What's the problem? Psychonomic Bulletin and Review, 22(2): 313-27.

Leising, Kenneth J., L. Caitlin Elmore, Jacquelyne J. Rivera, John F. Magnotti, Jeffrey S. Katz, and Anthony A. Wright (2013). Testing visual short-term memory of pigeons (*Columba livia*) and a rhesus monkey (*Macaca mulatta*) with a location change detection task. Animal Cognition, 16: 839-844.

Liebal, Katja, Simone Pika, Josep Call, and Michael Tomasello (2004). To Move or Not to Move: How Apes Adjust to the Attentional State of Others. Interaction Studies, 5(2): 199–219.

Lurz, Robert (2011). Mindreading Animals: The Debate over What Animals Know about Other Minds. Cambridge, MA: MIT Press.

Mayo, Deborah G. (1996). Error and the growth of experimental knowledge. Chicago: The University of Chicago Press.

Mayo, Deobrah G., & Spanos, Aris (Eds.) (2008). Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science. New York, NY: Cambridge University Press.

Meketa, Irina (2014). A critique of the principle of cognitive simplicity in comparative psychology. Biology & Philosophy, 29: 731-745.

Melis, Alicia P., Josep Call, and Michael Tomasello (2006). Chimpanzees (*Pan troglodytes*) Conceal Visual and Auditory Information from Others. Journal of Comparative Psychology, 120(2): 154–62.

Penn, Derek. C. (2011). How folk psychology ruined comparative psychology and what scrub jays can do about it. In R. Menzel, & J. Fischer (Eds.), Animal Thinking: Contemporary Issues in Comparative Cognition (pp. 253–265). Cambridge, MA: MIT Press.

Penn, Derek C., and Daniel J. Povinelli (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-Biological Sciences, 362(1480): 731-744.

Penn, Derek. C., & Povinelli, Daniel. J. (2009). On becoming approximately rational: The relational reinterpretation hypothesis. In S. Watanabe, A. P. Blaisdell, L. Huber, & A. Young (Eds.), Rational Animals, Irrational Humans. Tokyo, Japan: Keio University Press.

Penn, Derek. C., & Povinelli, Daniel. J. (2013). The comparative delusion: Beyond behavioristic and mentalistic explanations for nonhuman social cognition. In H. S. Terrace & J. Metcalfe (Eds.), Agency and joint attention. New York, NY: Oxford University Press.

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

Povinelli, Daniel J., Nelson, Kurt E., and Boysen, Sarah T. (1990). Inferences about guessing and knowing by chimpanzees (Pan troglodytes). Journal of Comparative Psychology, 104(3): 203-210.

Povinelli, Daniel J., and Timothy J. Eddy. (1996). "What Young Chimpanzees Know about Seeing." Monographs of the Society for Research in Child Development 61 (3): 1-152.

Povinelli, Derek. J., & Vonk, Jennifer. (2006). We don't need a microscope to explore the chimpanzee's mind. In S. Hurley, & M. Nudds (Eds.), Rational Animals? (pp. 385–412). New York, NY: Oxford University Press.

Premack, David, and Guy Woodruff (1978). “Does the Chimpanzee Have a Theory of Mind?” Behavioral and Brain Sciences 1(4): 515-26.

Sani, Fabio and John Todman. (2006). Experimental Design and Statistics for Psychology: A First Course. Oxford: Blackwell Publishing.

Shettleworth, Sara J (2010). Clever animals and killjoy explanations in comparative psychology. Trends in Cognitive Sciences, 14(11): 477-481.

Sober, Elliott (2001). The principle of conservatism in cognitive ethology. Royal Institute of Philosophy Supplement, 49: 225-238.

Staley, Kent (2008). Error-statistical elimination of alternative hypotheses. Synthese, 163: 397-408.

Tomasello, Michael, and Josep Call. (2006). “Do Chimpanzees Know What Others See—or Only What They Are Looking At?” In Rational Animals?, ed. Susan Hurley and Matthew Nudds, 371–84. Oxford: Oxford University Press.

Worrall, John (1989) Fresnel, Poisson and the white spot: The role of successful predictions in the acceptance of scientific theories. In David Gooding, Trevor Pinch, and Simon Schaffer (Eds.) The uses of experiment: Studies in the natural sciences. Cambridge, UK: Cambridge University Press.