Psychon Bull Rev
DOI 10.3758/s13423-014-0704-4

THEORETICAL REVIEW

Animal mindreading: what's the problem?

Cecilia Heyes

Abstract Research on mindreading in animals has the potential to address fundamental questions about the nature and origins of the human capacity to ascribe mental states, but it is a research programme that seems to be in trouble. Between 1978 and 2000 several groups used a range of methods, some with considerable promise, to ask whether animals can understand a variety of mental states. Since that time, many enthusiasts have become sceptics, empirical methods have become more limited, and it is no longer clear what research on animal mindreading is trying to find. In this article I suggest that the problems are theoretical and methodological: there is difficulty in conceptualising alternatives to 'full-blown' mindreading, and reluctance to use the kinds of empirical methods necessary to distinguish mindreading from other psychological mechanisms. I also suggest ways of tackling the theoretical and methodological problems that draw on recent studies of mindreading in humans, and the resources of experimental psychology more generally. In combination with the use of inanimate control stimuli, species that are unlikely to be capable of mindreading, and the 'goggles method', these approaches could restore both vigour and rigour to research on animal mindreading.

Keywords Animal cognition · Animal learning · Comparative cognition · Mentalising · Mindreading · Social cognition · Social understanding · Theory of mind

C. Heyes (✉)
All Souls College and Department of Experimental Psychology, University of Oxford, Oxford, UK
e-mail: cecilia.heyes@all-souls.ox.ac.uk

Published online: 08 August 2014

Introduction

The capacity to ascribe mental states, such as beliefs and desires, to oneself and to other agents is widely regarded as fundamental to human social life. Known as 'theory of mind', 'mentalising', 'folk psychology', 'social understanding' and 'mindreading', this capacity enables us—in everyday life and through law, politics and education—to predict, explain, justify and regulate the behaviour of others. Two important questions in current research involving infants, children and adults concern the origins and nature of mindreading: To what extent is the capacity for mindreading inborn, rather than inherited culturally or constructed through individual experience (Heyes & Frith 2014)? Is there an 'implicit' or 'automatic' form of mindreading, in addition to the 'explicit', deliberative processes that have been assumed traditionally to constitute thinking about mental states (Apperly 2011)? In addition to telling us a great deal about the evolution of cognition, research on mindreading in nonhuman animals (henceforth 'animals') could play a major role in resolving these debates. For example, clear evidence of mindreading in animals would challenge the view that human mindreading is inherited culturally, and support the proposal that human infants are capable of implicit mindreading. However, after some 35 years of research on mindreading in animals (Premack & Woodruff 1978), there is still nothing resembling a consensus about whether *any* animal can ascribe *any* mental state (Buckner 2013; Call & Tomasello 2008; Lurz 2011; Penn & Povinelli 2007, 2013; Whiten 2013).

In this article I argue that the lack of consensus in research on animal mindreading is due to an unusual set of problems. It is not just a matter of slow progress towards the resolution of a difficult and contentious scientific question. Rather, due to theoretical and methodological problems, the vigour and rigour of research on animal mindreading has declined in the last 15–20 years. To get back on track, the field needs a fuller and

clearer set of theoretical alternatives to 'full-blown' mindreading, and a return to the use of demanding but potentially effective empirical methods.

This article has three principal sections. The first offers a very brief history and overview of research on animal mindreading, highlighting the progress made in earlier years and the difficulties encountered more recently. The second discusses theoretical problems and how they might be resolved. In advance of that resolution, it is impossible to pinpoint experimental methods that would distinguish mindreading in animals from other psychological processes enabling the prediction of behaviour. However, the third section discusses empirical methods that may help to clear the log-jam in research on animal mindreading.

A history of animal mindreading in three experiments

Three experiments capture key developments in research on animal mindreading. The first (Woodruff & Premack 1979), along with the article in which Premack and Woodruff (1978) launched psychological research on theory of mind, began a period of diversification in which the primary objective was to find evidence that nonhuman primates could understand false belief. The second (Povinelli et al. 1990) represents a shift from false belief to seeing and knowing as targets of enquiry, and, I shall argue, a high point in methodological progress. The third (Hare et al. 2001) initiated the current era in which diversity has diminished and a large proportion of studies seek evidence that animals understand 'seeing' and 'knowing'. I have put these terms in scare quotes because, in recent years, researchers have emphasised that an animal's understanding of 'seeing' and 'knowing' may be very different from that of an adult human mindreader (see section below on Theory.) These three experiments also represent changes in core experimental methods, from conditional discrimination training alone, to conditional discrimination training followed by transfer tests, to the use of tests without prior training.

False belief / conditional discrimination training

Woodruff and Premack (1979) used conditional discrimination training to ask whether chimpanzees can learn to deceive; specifically, to act with the intention of inducing a person to hold a false belief about the location of food.[1] In each trial, the chimpanzee first saw food placed in one of two containers, both of which were out of reach. Then a human trainer entered the room, under instruction to search the container that the chimpanzee seemed to be indicating through its orientation or pointing behaviour. Sometimes the trainer was dressed in green and cooperative; if they found the food, they gave it to the chimpanzee. In other trials the trainer was dressed in white and competitive; if this trainer found the food, they kept it and the chimpanzee ended the trial empty handed. Analysis of the trainers' success rates in finding the food indicated that the chimpanzees learned the discrimination; to point towards the baited container in the presence of the cooperative trainer, and the empty container in the presence of the competitive trainer.

Subsequently it was recognised by Premack and others that the results of Woodruff and Premack's (1979) experiment did not provide evidence that chimpanzees are capable of intentional deception. Rather than seeking to induce a false belief in the competitive trainer, and a true belief in the cooperative trainer, the chimpanzees may have learned—associatively or via a more complex inference process—that pointing to the empty container was rewarded in the presence of the trainer dressed in white, and pointing to the baited container was rewarded in the presence of the trainer dressed in green. In other words, their pointing behaviour could have been based solely on thinking about 'observables'—stimuli that were physically present—rather than about mental states (Heyes 1998). However, in combination with Premack & Woodruff's (1978) inspiring discussion of theory of mind, Woodruff and Premack's experiment launched a thriving research enterprise. For about 20 years after these articles were published a number of research groups actively investigated animal mindreading using a range of field and laboratory methods. Much of the effort was devoted to finding evidence of intentional deception (Whiten & Byrne 1988), but many other facets of mindreading were also pursued in studies of imitation (Tomasello et al. 1993), self-recognition (Povinelli 1987), social relationships (Cheney et al. 1986), role-taking (Povinelli et al. 1992) and perspective-taking (Cheney & Seyfarth 1990; Povinelli et al. 1990) (see Heyes 1998 for a review).

Seeing / conditional discrimination training and transfer tests

The experiments on perspective-taking sought evidence that animals can represent what others see, and therefore what they know about the location of food. One of these attempted to avoid the ambiguity of Woodruff and Premack's (1979) results using a procedure in which chimpanzees were first given conditional discrimination training and then a transfer test (Povinelli et al. 1990). In each trial in the training phase, one of four cups was baited in the presence of the chimpanzee and a human trainer. Because this person saw the baiting, s/he was called the 'Knower'. The chimpanzee could see that the Knower was present, and that one of the cups was being baited, but not the exact location of the food. After baiting, a second trainer, the 'Guesser', entered the room, and each trainer pointed at a cup—the Knower at the baited cup, and the Guesser at one of the other three cups, chosen at random. The chimpanzee was then allowed to select one cup to search

[1] Woodruff and Premack (1979) also tested for comprehension of deceptive communications.

for food. If it selected the cup indicated by the Knower, it was allowed to eat the food, and if it selected the cup indicated by the Guesser or another cup, it ended the trial without reward.

Two chimpanzees mastered the discrimination; they learned to select the cup indicated by the Knower more often than the cup indicated by the Guesser. Because the roles of Knower and Guesser were assigned randomly to the two trainers in each trial, the chimpanzees could not have based their decisions on the trainers' appearance (cf. Woodruff & Premack 1979). Nonetheless, recognising that the successful chimpanzees could have been guided by another observable cue, which trainer was physically present during baiting, Povinelli et al. (1990) gave them a transfer test in which both trainers were in the room during baiting, but the Guesser had a bag over his head and therefore could not see where the food was placed. Across all 30 trials of the transfer test, the chimpanzees chose the cup indicated by the Knower more often than the cup indicated by Guesser, but subsequent analysis revealed that discrimination performance was at chance in the first five trials with the 'bagged' Guesser (Heyes 1993; Povinelli 1994).

This result was disappointing because it suggested that, rather than basing their choice responses on what they understood the trainers to have seen, and therefore to know, the chimpanzees had used two sets of observable cues. For example, in the training phase the chimpanzees may have learned to select the person who had been present during baiting, and in the transfer phase, the person who had been bag-free during baiting, without having any thoughts about *why* present, unbagged trainers provided reliable cues to the location of food. Ultimately, therefore, the experiment by Povinelli et al. (1990) did not provide evidence that animals understanding seeing or knowing. However, this study introduced to research on animal mindreading a 'triangulation' method—conditional discrimination training followed by transfer tests—that has long been recognised elsewhere in psychology as a powerful means of identifying what a subject knows (Campbell 1954). When an agent shows a systematic preference for one option over another, there are as many potential bases for this decision as there are detectable differences between the two options. Carefully chosen transfer tests enable researchers to isolate which perceptual inputs are contributing to the decision. In combination with theories specifying the kinds of input used by different psychological processes, this allows researchers to work out which psychological process mediated the decision. Thus, further experiments using the triangulation method could have been fruitful (Heyes 1998), but research on animal mindreading took a different turn.

'Seeing' / tests only

In 2001, Tomasello and his colleagues in Leipzig published a series of experiments seeking evidence that chimpanzees use an understanding of 'seeing' and 'knowing' to outflank opponents in a competitive feeding situation (Hare et al. 2001).[2] In these experiments, animals experienced different action–outcome relationships following different stimulus configurations, but the experiments were not intended to provide conditional discrimination training. Rather, it was hoped that the animals' choice responses under different stimulus conditions would reveal cognitive capacities that had developed, by some unspecified means, before the experiments began. In this sense, the competitive feeding paradigm, which has been used in many subsequent studies, involves testing without prior training.

At the beginning of each trial in the competitive feeding paradigm, a subordinate chimpanzee (the subject) and a dominant chimpanzee (the putative target of mindreading) were confined on opposite sides of an enclosure containing two occluding objects (Hare et al. 2001; Fig. 1). In all trials, a human trainer entered the enclosure and placed food on the subordinate's side of one of the occluders, and in some trials the trainer re-entered the enclosure 5–10 s later and moved the food to the subordinate's side of the other occluder. In all conditions, the door to the subordinate's cage was open during the baiting event(s). The conditions varied according to whether the dominant's door was open or closed, and therefore whether the subordinate could see the dominant, during the baiting event(s). After baiting, both of the chimpanzees were released into the enclosure, with the subordinate being given a head start.

The results suggested[3] that subordinates were more likely to secure the food, and less likely to refrain from approaching it, when (1) the dominant's door was closed rather than open during trials with a single baiting event, and (2) in trials where there were two baiting events and the dominant's door, although open during the first, was closed during the second baiting event. Furthermore, a follow-up experiment (Hare et al. 2001, Experiment 2) indicated that, in trials where there was a single baiting event with the dominant's door open, subordinates were more likely to get the food when they competed at the end of the trial with a different dominant individual than the one who witnessed baiting.[4]

The Leipzig group interprets these results, and those of other experiments using the competitive feeding paradigm (e.g. Bräuer et al. 2007; Hare et al. 2000), as evidence that chimpanzees have some understanding of the relationship between 'seeing' and 'knowing' (Call & Tomasello 2008). On this account, subordinates are less likely to approach and obtain the food when the

[2] The first studies using the competitive feeding paradigm were published in the previous year (Hare et al. 2000). I have focussed on the article published in 2001 because it reported procedures that have been used in many subsequent experiments.

[3] The published report (Hare et al. 2001) does not indicate clearly which of the contrasts were statistically significant because the conditions were not labelled consistently in the methods and results sections of the report.

[4] A third experiment reported by Hare et al. (2001) is not discussed here because it had null results.

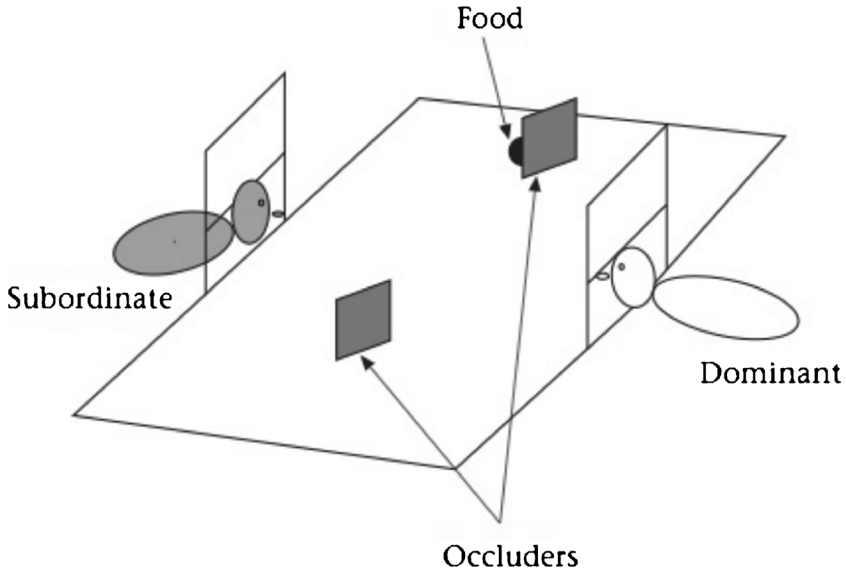

Fig. 1 Experimental set-up used with chimpanzees in the original competitive feeding paradigm. Reprinted from Hare et al. (2001)

dominant competitor's door was open because the subordinates understand that the dominant 'saw' food placement, and will therefore 'know' where the food is located. The subordinates assume that the dominants will use their knowledge, along with their dominant status, to win the feeding competition, and therefore do not try so hard. In contrast, other prominent researchers in the field—who were once enthusiastic believers in chimpanzee mindreading—are sceptical (Penn & Povinelli 2007; Whiten 2013). Povinelli and his collaborators in Louisiana have reported difficulties in replicating results from the competitive feeding paradigm (Karin-D'Arcy & Povinelli 2002). At minimum, these difficulties suggest that the published reports of the Leipzig experiments have not contained sufficient detail to allow their methods to be reproduced in other laboratories. Given more weight, these difficulties suggest that the competitive feeding paradigm does not exert sufficient control over the dominants' behaviour to exclude the possibility that the results reported by the Leipzig group are driven by the knowledge-based behaviour of the dominant chimpanzees, not by the knowledge (or any other mental attribute) ascribed to the dominants by the subordinates. A dominant that knows where the food was hidden, because their door was open, is likely to get to the food faster than a dominant that does not know where the food was hidden, and therefore more effectively to discourage the subordinate from approaching and securing the food (cf. Schmelz et al. 2011, 2013).[5]

[5] The Leipzig group has recently used a more carefully controlled "back-and-forth" competitive feeding paradigm in an attempt to rebut this "evil eye" hypothesis (Schmelz et al. 2011, 2013). The methods and results have not been reported in sufficient detail to establish whether the back-and-forth procedure has been successful in this respect. Even if it has, the possibility remains that the choice behaviour of subject chimpanzees in these studies depends, not on the ascription of preferences to others, but on learning – during or before the experiment—that inhibition of a prepotent response is rewarded under delayed test conditions.

Even if one takes the results from the competitive feeding experiments at face value, putting aside concerns about replicability and control of the dominants' behaviour, they do not overcome the 'observables' problem (Heyes 1998; Penn & Povinelli 2007; Whiten 2013). For example, a subordinate may be more hesitant to approach the food when their competitor was visible during baiting because they know that *visibility-during-baiting* predicts that the competitor will get the food. Regardless of *how* they know about this predictive relationship—via inferential or associative learning, before or during the experiment, or even if the knowledge is to some degree inborn—this knowledge is potentially sufficient to explain the subordinates' behaviour. They could, but they need not, 'understand' or have a 'theory' about *why* the relationship holds; they need not explain it to themselves with reference to what the dominant has 'seen' and therefore 'knows'.

A recent review of research on animal mindreading since 2000 noted that, although a wider range of species have been tested than in previous years—including dogs, elephants, pigs, and birds of the corvid family—the field has been dominated by studies of apes conducted by the Leipzig group using variants of competitive feeding paradigm (Whiten 2013). This review also implied—in common with other recent commentators (Buckner 2013; Lurz 2011) and my own view—that 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.

Overview

This brief history of research on animal mindreading illustrates three important trends. First, the bar has been lowered.

Researchers initially sought evidence that animals can represent false beliefs; then targeted human-like understanding of seeing and knowing; and are now typically asking whether animals have *any* understanding of 'seeing' and 'knowing'. Second, viewed from the perspective of experimental psychology, methodological standards have declined. In the last 15 years or so, the methods and results of experiments on animal mindreading have often been reported with less precision than in previous years; alternative explanations for the data abound but are rarely given careful consideration in empirical papers; and many of the experimental designs—using tests without prior, formal training—lack the potential to identify the observable cues and cognitive processes used by animals to make social decisions. Third, the social structure of research on animal mindreading has changed. In earlier years there were a number of active research groups, each publishing a significant volume of empirical work and voicing their own theoretical perspectives. More recently, although groups studying dogs and corvids have emerged (Clayton et al. 2007; Topal et al. 2009), the field has been dominated by the Leipzig group, and previously influential and enthusiastic contributors no longer do empirical work on animal mindreading. In their commentary articles and reviews, these researchers now express doubts (Seyfarth & Cheney 2012; Whiten 2013) or outright scepticism (Penn & Povinelli 2007, 2013). They stress the difficulty of finding out whether animals are capable of mindreading, and in some cases suggest simply that animals do not read minds.

So, in these respects, research on animal mindreading has declined. Why? I suggest that the underlying problems are theoretical and methodological. It is no longer clear what research on animal mindreading is looking for (theoretical problem), and consequently it is not clear how the quarry can be hunted down (methodological problem).[6]

Theory

The problem

There is no proprietary definition of what it is to engage in mindreading or to have a theory of mind. The term 'theory of mind' embodies a broad claim about the nature of our (adult human) understanding of mental states—that it is theoretical, or comprises a "system of inferences" (Premack & Woodruff 1978)—but the type and content of these inferences has been a topic of philosophical debate for centuries, and even the broad claim has been challenged by simulationists, past and present (Goldman 2006). Consequently, researchers who want to find out whether animals are capable of mindreading have considerable freedom in how they define animal mindreading. However, I suggest that, to avoid miscommunication and allow empirical progress, there are at least three important constraints, the 'three Es':

(1) *Eccentricity*. A definition—or, more loosely, conception—of animal mindreading that departs significantly from common, contemporary usage of the term in psychology and philosophy is at risk of being misunderstood. Therefore, a helpful conception of animal mindreading is likely to stay close to a currently conventional view of what it is to understand mental states, or to explain and justify its innovative features very clearly. Thus, the first E constraint can be met either by avoiding or by carefully justifying eccentricity.

(2) *Evolutionary precursors*. Not all research on animal cognition is anthropocentric but the primary purpose of research on animal mindreading is to find evolutionary precursors of human mindreading. Therefore, a useful conception of animal mindreading will facilitate this task.

(3) *Empirical testing*. Psychological, as opposed to philosophical, research on animal mindreading tries to make progress through observation and experiment. Therefore, a helpful conception of animal mindreading will provide a basis for the formulation of target and alternative hypotheses; of empirically testable proposals about what is, and is not, going on in an animal's head when it is mindreading.

In the light of these constraints, it was entirely reasonable for the Leipzig group to question the conception of animal mindreading that guided research until around 2000. Although never fully spelled out, at its richest the earlier conception assumed that animal mindreading would involve metarepresentation of propositional attitudes, including false beliefs. For example, mental representation of a mental representation such as "He believes [attitude] the food is in the red container [proposition]", when the mindreader believes that the food is in the blue container. Such a "full-blown" (Butterfill & Apperly 2013) conception of mindreading may well be setting the bar too high for animals—or indeed for any agents, including human infants, who do not have language. Specifically, although it meets the eccentricity constraint by conforming to a demanding but commonly held view of adult human mindreading, the full-blown conception is, precisely because it meets the eccentricity constraint in this way, unlikely to reveal evolutionary *precursors* of human mindreading—steps on the way, or on a path similar to, the evolutionary trajectory that led to mindreading in humans.

[6] The situation may be exacerbated by a socioeconomic problem: due to changes in the social structure of research on animal mindreading, the theoretical and methodological problems are less likely to be solved because fewer minds are working on them, and new ideas are less likely to be honed by disagreement.

However, in my view the "perception-goal psychology" conception of animal mindreading, with which the Leipzig group replaced the full-blown conception, sets the bar too low. It suggests, for example, that animal A understands what another animal, B, 'sees', if A understands "not just what he [B] is oriented to, but what he [B] registers from the environment in ways that affect his actions" (Call & Tomasello 2008, p 189). Thus, 'seeing' is contrasted with orienting and identified with 'registering'. This is a promising start, but perception-goal psychology does not tell us enough about 'registering'. What are the conditions in which A understands B to have registered an event, and how do they differ from those in which A understands B merely to have oriented to an event? How does A mentally represent registration, rather than orientation, by B? How does A come to be able to represent registration by B?[7] Without addressing these questions, the perception-goal psychology conception of mindreading is at risk of failing to meet all three E constraints: (1) it is eccentric—previous conceptions of mindreading did not cast 'registering' as a mental state—and its innovative features are not explained and justified. (2) It is unlikely to support the discovery of specific evolutionary precursors of human mindreading because, without a clear account of how understanding registering differs from understanding orienting, it could be argued that all animals that are sensitive to the orientations of other animals' bodies are also capable of registering and therefore of mindreading. Given that sensitivity to body orientation is necessary for most forms of social interaction, this would make a very broad range of animals into mindreaders. The capacity would not be confined to taxa that are closely related to humans (e.g. primates), or that plausibly represent a convergent path of evolution (e.g. birds). (3) Clear conceptual distinctions are the bedrock of empirically testable hypotheses. Therefore, in the absence of a clear conceptual distinction between understanding registering and understanding orienting it is not possible to formulate testable hypotheses about what an animal will do in an experimental situation if it understands registering as well as orienting (target hypothesis), rather than orienting alone (alternative hypothesis).

In summary, I am suggesting that the core theoretical problem in contemporary research on animal mindreading is that the bar—the conception of mindreading that dominates the field—is too low, or more specifically, that it is too underspecified to allow effective communication among researchers, and reliable identification of evolutionary precursors of human mindreading through observation and experiment. Two potential solutions to this theoretical problem have recently been proposed: the 'intervening variable' solution, and the 'minimal' solution.[8]

The intervening variable solution

Whiten (2013) has recently endorsed a potential solution to the theoretical problem that he first proposed some 20 years ago (Whiten 1994). His intervening variable conception of animal mindreading is highly reflexive; it suggests that an animal mindreader is like a comparative psychologist who is trying to come up with an economical model of animal behaviour. Just as the comparative psychologist postulates an intervening variable such as 'thirst' to explain the modulatory effects of hours of water deprivation, consumption of dry food, and saline injections (environmental inputs) on rate of lever pressing for fluid, volume of fluid consumed, and quinine tolerance (behavioural outputs), the mindreading animal postulates an intervening variable such as 'B knows food is in X' to explain the modulatory effects of various environmental inputs on animal B's social behaviour (Fig. 2).

The reflexivity of the intervening variable conception of mindreading makes it very appealing, at least for comparative psychologists. We have had lots of practice in trying to explain animal behaviour with reference to intervening variables. Consequently, we know at some level what that endeavour involves for us, and therefore, according to the intervening variable solution, we know what mindreading would involve for an animal. Thus, the intervening variable solution seems to meet the eccentricity constraint with flying colours by avoiding eccentricity, and to give us a clear target for empirical enquiry. However, this appealing feature also inclines the intervening variable solution to set the bar for animal mindreading too high for the identification of evolutionary precursors of human mindreading. If animal mindreading is just like what comparative psychologists are doing, there's a risk that it involves full-blown theory of mind. Indeed, Fig. 2, which is reproduced from Whiten's articles (2013), implies that animal mindreading involves metarepresentation of propositional attitudes. The central box in the box-and-arrow diagram, depicting the intervening variable used by A to explain B's behaviour, seems to be a mental representation of B's attitude (knowing) towards a proposition (the food is in X).

It could be argued with some force that there is still a place in research on animal mindreading for a user-friendly conception of full-blown theory of mind of the kind offered by the

[7] One could also ask what is meant by 'understanding', but this uncertainty is not specific to perception-goal psychology. It is now commonplace in research on animal and infant cognition to discuss what participants 'understand' without explaining what this term implies.

[8] Buckner (2013) argues persuasively that the core problem in contemporary research on animal mindreading is semantic; the field is concerned with 'concepts' and 'representations' of mental states, entities that are defined by their content, and yet researchers—both enthusiasts and skeptics—do not specify or defend their assumptions concerning how we should decide what a representation is about. On this view, both the intervening variable and minimal mindreading solutions are potential answers to the semantic problem.

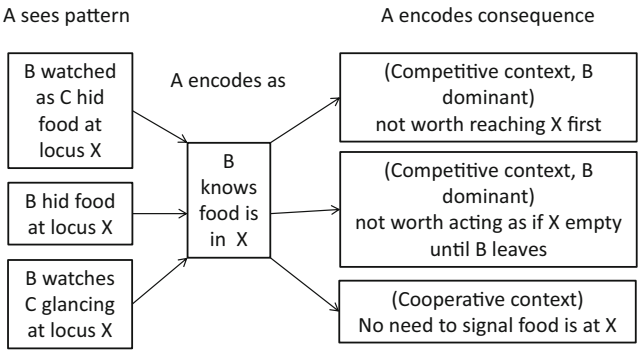

Fig. 2 The diagram used by Whiten to represent his intervening variable account of animal mindreading. Reprinted from Whiten (2013)

intervening variable solution. In my view and Whiten's (2013), animals have not yet been tested systematically for mindreading using the most promising methods, and therefore it may be too soon to give up on finding that some animals have full-blown theory of mind. However, as it is currently formulated, the intervening variable conception has limitations with respect not only to the second but also to the third E constraint; it does not provide a firm basis for the formulation of testable target and alternative hypotheses.

Whiten (2013) contrasts intervening variable ascription with learning "pairwise links", and suggests that these target and alternative hypotheses can be empirically distinguished by testing whether novel inputs (inputs are on the left side of Fig. 2) "drive the same adaptive outputs" (right side of Fig. 2) as familiar inputs. This implies that experimental work could be guided by a target, mindreading hypothesis proposing that the behaviour of subject animals (As) is controlled by mental representations of the kind depicted in Fig. 2, contrasted with an alternative, non-mindreading hypothesis in which it is controlled by associative links between each of the inputs on the left of Fig. 2 and each of the outputs on the right, and that these hypotheses could be distinguished by stimulus transfer tests.

To see how this might work in practice let us consider a hypothetical case in which we have found a chimpanzee, Arthur, who seems to know about nearly all of the relations represented in Fig. 2; the relations involving the top two boxes on the left and all three boxes on the right. Thus, when Arthur has seen another animal, B, hiding food at X (the 'B hiding' stimulus), and when Arthur has seen B watching food being hidden at X (the 'C hiding' stimulus), if B is a dominant competitor, Arthur is slower to approach X, and if B is his friend, Arthur is less likely to make a food call at X. Now we do a transfer test in which we allow Arthur to see B oriented towards a third party, C, as C repeatedly looks at location X (the 'glancing' stimulus described in the bottom left hand box of Fig. 2). He may have seen this 'glancing' stimulus before but, crucially, we have it on good authority that the 'glancing' stimulus is "novel" in that Arthur has never, after exposure to this stimulus, tried and failed to get food at X when B was a dominant competitor. If in the transfer test Arthur responds to 'glancing' by slowing his approach to X, and/or by inhibiting a food call, then according to Whiten's proposal we would have evidence in support of the target hypothesis that Arthur is ascribing knowledge to B, and against the alternative view that he has merely learned "pairwise links" connecting the left and right boxes in Fig. 2.

However, that inference would not be valid because there are at least two ways in which pairwise learning could produce Arthur's successful transfer test performance. First, stimulus generalisation: Having based his initial performance on pairwise links, Arthur may have responded to the 'glancing' stimulus as if it was the 'B hiding' or 'C hiding' stimulus because the 'glancing' stimulus physically resembled the hiding stimuli. Second, mediated conditioning: an animal that hides food at X is likely to glance at X before and after doing so. Therefore, it is likely that Arthur had experienced pairings between the 'glancing' and 'C hiding' stimuli, or that each of these stimuli had been paired in his experience with a common outcome, such as frustration at being unable to access the food. In both of these cases, experiments on sensory preconditioning (Rizley & Rescorla 1972) and mediated conditioning (Hall 1991; Honey & Hall 1989) in rats suggest that the 'glancing' and 'C hiding' stimuli would become associated with each other such that presentation of one of these stimuli would activate representations of them both, and thereby allow "pairwise" learning (links across Fig. 2) involving the stimulus that was not physically present. Consequently, although Arthur has not had the kind of *direct* experience that would allow him to learn that approach to X is not rewarded following the 'glancing' stimulus, he could have learned this kind of pairwise link *indirectly*, when exposure to the 'C hiding' stimulus activated a representation of 'glancing'.

This example suggests that, at least as it is currently formulated, the intervening variable solution falls short of providing a firm basis for empirical enquiry about animal mindreading because—in common with most contemporary research in this area—it does not take seriously the formulation of alternative hypotheses. Research on associative learning is, as the allusion to "pairwise links" implies, one potential source of such hypotheses, but to fulfil this potential it needs to be mined more thoroughly (Heyes 2012). This is not an attractive prospect for most researchers because the literature on associative learning is highly technical. But—much as I would like to do it myself sometimes—this literature cannot be ignored simply because it is demanding. Until research on animal mindreading makes use of explanatory resources from contemporary associative learning theory, and elsewhere in psychology and cognitive science (see below), it will remain at high risk of reaching false positive conclusions.

In summary, I have argued that, as it has been presented by Whiten (2013), the intervening variable conception of animal mindreading is strong with respect to the eccentricity

constraint, and defensible in relation to evolutionary precursors, but—on close examination—weak as a basis for the formulation of empirically testable target and alternative hypotheses.

The minimal solution

In my view, Butterfill and Apperly's (2013) "minimal theory of mind" provides an exceptionally promising potential solution to the theoretical problem at the heart of contemporary research on animal mindreading. It offers a conception of mindreading—applicable to infants, adults and animals—that lowers the bar relative to the full-blown conception, but does so in a principled and clearly articulated way. The 'minimal solution' raises the possibility that, rather than involving metarepresentation of propositional attitudes, animal mindreading involves the representation of mental states as relations between agents, objects and locations. According to this view, animals do not represent mental states as such, but their representations of certain relations count as a variety of mindreading because, Butterfill and Apperly argue, they "track" mental states. In other words, it is facts about mental states that make these representations successful in controlling social behaviour. To use Butterfill and Apperly's analogy, the minimal solution suggests that an animal's understanding of mental states may be like an animal's understanding of toxicity. It is very unlikely that any animal knows the things known by a human toxicologist—about routes of exposure, dose-response curves, and *why* certain chemicals have adverse effects on living organisms. However, an animal's representations of odours and visual stimuli associated with putrefaction could yield nutritional and competitive benefits by virtue of facts about toxicity.

Butterfill and Apperly characterise minimal theory of mind using four "principles". Loosely speaking, these are four things that the possessor of minimal theory of mind understands: (1) goal-directedness—some bodily movements are performed because they have had a certain outcome in the past (goal-directed movements) while others are not. 2) Encountering—goal-directed action on an object requires that the object has been 'encountered', where 'encountering' depends on the object being in the agent's 'field', and what is in the agent's field depends on spatial and physical constraints such as proximity and lighting, as well as the agent's orientation and posture. Thus, encountering is like perception, but an animal that understands encountering need not understand it to be related to reasons or knowledge. (3) Registering—*successful* goal-directed action requires 'registration' of an object at a location, where an agent registers an object at a location only if the agent most recently encountered the object at that location. (4) Registering is causal—when an agent performs a goal-directed action with a goal that specifies a particular object, she will act as if the object is at the location where she registered the object. In combination, principles 3 and 4 make 'registrations' somewhat like beliefs. They generalise across all goal-directed actions, can be assigned correctness conditions, and they causally influence action. However, unlike beliefs, registrations are not propositional attitudes, they cannot refer to objects that have never existed, and they are not subject to the norms that characterise beliefs (Butterfill & Apperly 2013).

The great strengths of this minimal solution lie in the way it meets the eccentricity and evolutionary precursor constraints on a conception of animal mindreading. Like perception-goal psychology, minimal theory of mind is eccentric in that it departs significantly from the full-blown conception, but Butterfill and Apperly explain and justify the departures with great care. For example, in contrast with perception-goal psychology, which does not adequately distinguish registering from orienting (see above), minimal theory of mind characterises registration as a belief-like state with a specified functional role in relation to goal-directed action, and identifies orientation as a behavioural variable that contributes to determining whether a particular object was in an agent's field, and therefore whether it was encountered. Butterfill and Apperly also justify the eccentricity of the minimal solution by showing how the innovative features of minimal theory of mind make it possible to conceptualise a limited form of mindreading of which animals (and infants, and adults under cognitive pressure) may be capable. They do not cast minimal theory of mind as an evolutionary or developmental 'precursor' in the sense of being a necessary step in the phylogenesis or ontogenesis of full-blown mindreading, but it represents a plausible platform for both of these. So, in my Goldilocks view, minimal theory of mind does not set the bar for animal mindreading too high or too low; in terms of conceptual level, it is just right.

My reservations about the minimal solution relate to the third E constraint—the formulation of empirically testable target and alternative hypotheses. Butterfill and Apperly (2013) are acutely aware of the importance of this constraint, and devote a significant proportion of their recent paper to discussing how "signature limits" might be used to distinguish empirically between minimal theory of mind and, on the one hand, full-blown mindreading, and on the other, the use of "behavioural strategies". I may have contributed to the emergence of this now conventional way of conceptualising alternatives to mindreading (Heyes 1998)—as behavioural strategies, "behaviour rules" or "behaviour reading"—but I think it is problematic.

Behaviour reading

Premack and Woodruff (1978) suggested that if animals are not "mentalists", they are "behaviourists". Twenty years later, I argued that, although deliciously witty, this contrast had been misleading (Heyes 1998). It had encouraged researchers to use behaviourism, or stimulus–response (S–R) learning theory, as

their only source of alternative hypotheses, when the set of potential alternatives is in fact much larger. I argued that the debate about animal mindreading is concerned primarily with *what* animals represent—only observable features of other animals' situations and behaviour, or these features plus mental states—rather than *how* these contents are represented, for example, in an abstract or concrete way, via inferential or associative processes. Animals that represent only situations and behaviour could do so in an abstract, inferential way, but they still would not be mindreaders. My purpose in advancing this argument was to encourage researchers to look beyond S–R learning theory, to contemporary associative learning theory and other areas of psychology and cognitive science for testable alternative hypotheses. But that was not what happened. A broader conception of 'not mindreading' gained currency—labelled 'behaviour reading'—but this conception was and is based on common sense rather than cognitive science.

Any conditional statement that a researcher can imagine, referring to behaviour and not to mental states, currently counts as a behavioural rule or strategy. Sometimes these statements are expressed in an elegant and semi-formal way. For example, Perner (2010) notes that successful performance in some complex mindreading tasks could be mediated by the behaviour rule: "If a person P looks at an object O being put inside a location L1, and does not look when it is transferred to location L2, and if P is to get the object O, then P will go to L1" (Perner 2010, p 249). But this style of presentation should not obscure the fact that the vast majority of behaviour rules considered in current research on mindreading are based on common sense categories (person, object, location), and are not supported or constrained by empirical evidence of any sort. If 'folk psychology' is our pre-scientific understanding of psychology, behaviour rules are as much a part of folk psychology as conditional statements referring to mental states.

Conceptualising the alternative to animal mindreading as behaviour reading generates three problems. The first is widely recognised: because behavioural strategies are so unconstrained—limited by imagination rather than evidence—it is very difficult indeed, perhaps impossible, to design experiments that could show that animals are mindreading rather than behaviour reading. How, for example, could one secure solid evidence that an animal (or pre-linguistic infant) has ascribed a false belief about an object's location, rather than applied Perner's (2010) behaviour rule, quoted in the previous paragraph? Some authors have gone so far as to suggest that this is a "logical" (Hurley & Nudds 2006; Povinelli & Vonk 2004) or "Laplace's demon" (Butterfill & Apperly 2013) problem, implying that all putative, nonverbal evidence of mindreading could in principle be explained by behavioural rules. This may be correct but, if so, it is surprising that the same authors go on to suggest empirical methods (discussed below) that might overcome the problem. Surely, if the problem is "logical" or Laplacean, it cannot be solved by clever experimental methods, and if it can be solved by such methods, the problem is not logical but methodological or, as I am suggesting, primarily theoretical—due to over-reliance on common sense as a source of alternative hypotheses, and soluble by increasing reliance on cognitive science.[9]

Thus, the first problem with the concept of behaviour reading is that it encourages researchers to entertain *too many* alternative hypotheses; to take seriously behavioural rules for which there is no empirical support. Conversely—and this is the second problem—the concept of behaviour reading also encourages us to entertain *too few* alternative hypotheses; to overlook alternatives that have empirical support from cognitive science, but which are not evident to common sense. Elsewhere I have called these "submentalising" alternatives (Heyes 2014a, 2014b). These alternatives suggest that behaviour that seems to indicate mindreading is supported by much lower-level, domain-general processes. Unlike behavioural strategies, submentalising processes do not involve what most psychologists would describe as reasoning, and they parse the world in terms of features such as colour, shape and movement, rather than actions on objects by agents.

Submentalising alternatives are routinely neglected, even in work that is otherwise of exceptionally high quality (e.g. Bugnyar 2011; Samson et al. 2010). In one example, which Butterfill and Apperly (2013) cite as evidence that mature humans sometimes use minimal theory of mind rather than full-blown mindreading or behavioural strategies, Samson et al. (2010) asked adults to make speeded judgements about images showing an avatar—a human-like figure—standing in the centre of a room, facing to the right or to the left. There were dots on the walls of the room, in front of the avatar, behind the avatar, or both. When the participant was asked to judge the number of dots she could see (i.e. the total number of dots on the screen), responses were slower and less accurate in "inconsistent trials", where the avatar could see fewer dots than the participant (e.g. there was one dot in front of the avatar and one behind), than in "consistent trials", where the avatar could see the same number of dots as the participant (e.g. two dots in front of the avatar and none behind). There is compelling evidence that this consistency effect does not depend on executive functions (Qureshi et al. 2010), and is therefore unlikely to be due to full-blown representation of what the avatar can see, or to an incorrect behavioural rule, 'report the number in front of the avatar', interfering with the correct behavioural rule, 'report the number on the screen'. However, if one looks to cognitive science rather than common sense, this is not sufficient to show

[9] Buckner (2013) suggests that whether or not there is a distinctive logical problem at the heart of research on mindreading depends on one's theory of representation, and that researchers seldom make explicit their theories of representation. If this is correct, it raises the possibility that researchers sometimes express contradictory views—that the problem is logical *and* that it can be solved by clever experiments – because they do not consistently apply a single, implicit theory of representation.

that the consistency effect is due to minimal theory of mind—to representation of what the avatar is "registering". Research on automatic attentional orienting (e.g. Tipples 2002, 2008) raises the possibility that the consistency effect is due to the directional rather than the agentive features of the avatar stimulus; that the 'front features' of the avatar (nose, chest, toes), like the point of an arrow, direct attention to the dots on one side of the avatar, and counting or subitising those dots interferes with the explicit task of judging the total number of dots on the screen. I have chosen this example, in spite of the fact that it involves humans rather than animals, because it is unusual in that the alternative hypothesis from cognitive science has been identified, tested against the minimal mindreading hypothesis, and received empirical support: the consistency effect is just as strong when the central stimulus is an arrow—an object that elicits automatic attentional orienting, but is not capable of 'registration'—as when the central stimulus is an avatar (Santiesteban et al. 2013).

The third and final problem with casting the alternative to mindreading as behaviour reading is more specific: minimal theory of mind seems to be species of, rather than distinct from, behaviour reading. The key constructs in minimal theory of mind—'field', 'encountering', 'registering'—are all defined by observable relations between agents, objects and locations, and it would appear that the four principles of minimal mindreading could be re-expressed in the kind of conditional statements typically used to specify behavioural rules. If this is correct, and if the concept of behaviour reading was not problematic in other ways, one might conclude that minimal theory of mind needs to be changed so that it becomes more distinct from behaviour reading. However, given that minimal theory of mind offers such a promising and long-awaited solution to the theoretical problem of animal mindreading, and that conceptualising the alternative to mindreading as behaviour reading generates two other significant problems, I think the third E constraint—empirical testing—would be better served by changing how we think about alternatives to mindreading. The minimal solution will no doubt benefit from further theoretical development but, in my view, revisions should not be made specifically in order to preserve the territory of behaviour reading.

The inconvenient implication of this discussion is that a better conception of 'not mindreading' would be more disparate and less dependent on common sense than the current conception of behaviour reading. Of course it would not exclude common sense. Very few areas of science could do that without disastrous consequences. But it would encompass (1) only those common sense behaviour rules for which there is, or could be, empirical support; and (2) a wide range of low-level, domain-general submentalising processes revealed by cognitive science. In my view, alternatives to mindreading should *not* be sought exclusively, or even especially, in associative learning theory; in the body of work, based on conditioning experiments, that casts learning as the formation of excitatory and inhibitory links between lean representations of events (Heyes 1998; Papineau & Heyes 2006). I drew on the associative tradition when discussing Whiten's intervening variable solution (see above) partly because Whiten himself invoked that tradition by referring to "pairwise links", and partly because I regard the associative tradition as a fine example of rigorous, cumulative, empirically based psychological research, which can explain a significant proportion (but not all) of the learning phenomena observed in a wide range of species, including humans (Heyes 2012; cf. Penn & Povinelli 2013). However, as illustrated by my discussion of automatic attentional orienting (above), encoding specificity (below), and retroactive interference (elsewhere, Heyes 2014a), there are many other areas of cognitive science—concerned with perception, attention and memory, as well as learning—that yield testable alternatives to mindreading hypotheses. The key feature of these sources of submentalising hypotheses is that their theoretical constructs have been honed, not by common sense, but by careful experimental investigation of robust behavioural and neurological effects.

The idea that alternatives to mindreading can be encapsulated in one homogeneous category, with a complementary name—behaviour reading—is dangerously appealing, but if it were true, research on animal mindreading would be very unusual. In research on animal navigation, for example, the question 'Do animals have cognitive maps?' has not been addressed by conceptualising the alternatives in terms of one homogeneous category. Rather, it is recognised that, instead or in addition to using cognitive maps for navigation, animals can use pheromone trails, dead reckoning, beacons, piloting with multiple cues, geometric relations and a host of other processes, none of which were discovered purely through common sense (Pearce 2008).

In summary, I have argued that the core problem in research on animal mindreading in the last 15 years has been theoretical; research has not been guided by a conception of animal mindreading that meets the three E constraints by either avoiding or justifying eccentricity; being apt for the discovery of evolutionary precursors of human mindreading; and enabling the formulation of empirically testable target and alternative hypotheses. In considering potential solutions to this problem, I have suggested that, whereas the intervening variable conception of animal mindreading meets only the first constraint, the minimal solution both justifies its own eccentricity, and provides a very promising 'search image' for evolutionary precursors of full-blown human mindreading. In combination with a new conceptualisation of alternatives to mindreading—which draws more heavily on cognitive science than on common sense—the minimal solution could also enable the formulation of empirically testable target and alternative hypotheses about animal mindreading.

Methods

Around 15–20 years ago the primary problem in research on animal mindreading was methodological. Researchers were more or less agreed about what they were looking for—metarepresentation of at least some propositional attitudes—but were in dispute about the methods required to test for this capacity (Heyes 1998). The primary problem is now theoretical. Researchers are not clear about what they are seeking, or how their quarry differs from other processes that could give the appearance of mindreading in animals. This situation generates methodological problems downstream—when you don't know what you're looking for, you don't know how to find it—but these can be resolved only after, or in concert with, progress in tackling the theoretical problem. Therefore, it is not possible at this stage to pinpoint methods that are likely to play a crucial role in future research on animal mindreading. However, using a final empirical case study, I will highlight some of the methodological implications of the theoretical points discussed in the previous section.

Building on the experimental designs presented in the first section of this article, Bugnyar (2011) recently tested for "knower–guesser differentiation" in ravens. I have chosen Bugnyar's experiments as a final case study because they are typical of post-2000 research in having used a variant of the competitive feeding paradigm (see 'Seeing' / tests only' above) and corvids as subjects,[10] but extraordinary in the care and precision with which they were designed, interpreted and reported. In my view, this is the best experimental work on animal mindreading published since 1990 (see 'Seeing / conditional discrimination training and transfer tests' above). The queries I will raise reflect problems, not with this study in particular, but with the state of the art in research on animal mindreading.

Each trial involved three ravens—the subject (S), who was the putative mindreader, and two observers (O1 and O2), who were in competition with the S for food. In the first, caching phase of each trial the experimenter buried cheese successively at each of two locations in the central enclosure, cache 1 and cache 2 (Fig. 3a). When cache 1 was made, O1 was visible to S on the other side of the enclosure, and when cache 2 was made, O2 was visible to S on the other side of the enclosure.[11] In Experiment 1, the visible observer was standing on the ground, as shown in Fig. 3a. In Experiment 2, the visible observer was standing on a perch 1.5 m above the ground, and in half of the trials, a curtain with a window was drawn down over the observer's cage (Fig. 3c). In these 'window trials', the S could see the observer but the lower part of the curtain obstructed the observer's view of where the food was cached. In the second, retrieval phase of each trial, the S was released into the central enclosure while O1 or O2 was confined but visible, standing on the ground, on the opposite side of the enclosure (Fig. 3b). Once the S had chosen a cache by touching the cover and/or retrieving the food, the visible bird was released, allowing it to compete with the S for the other cache after the performance measures had been taken. The results were clear cut: in 'no window' trials in both experiments the S ravens showed a significant tendency to match caches with competitors: they chose cache 1 when released in the presence of the competitor who had been visible when cache 1 was made, and cache 2 in the presence of the competitor who had been visible when cache 2 was made. In contrast, in window trials there was no systematic relationship between the identity of the competitor (O1 or O2) and the first cache chosen by the S.

In his sober and scholarly discussion of these results, Bugnyar (2011) argued:

(1) *Not behaviour reading*. The effect of the window manipulation excluded the possibility that the cache choice of the S ravens was based on a behaviour rule such as "compete with those that could be seen at the time of caching" or, as it was described elsewhere in the article, by learning "to associate specific competitors with specific cache sites".

(2) *Minimal mindreading*. Therefore, the ravens must have used "lines of sight" or 'eye-object line' (Heyes 1998). For example, when confronted by O1 in the retrieval phase, the S raven's choice must have been based on a memory of whether, in that trial, there had been an unobstructed straight line between the body of O1 and the first caching event or the second caching event.

(3) *Full-blown mindreading*. The S ravens may have used eye-object line as an indicator of what other ravens had seen and therefore knew. That is, ravens may be capable of "positing abstract 'intervening variables' that code for (some of) the others' mental states such as 'seeing' or 'knowing'" (p 639).

I have labelled these points 'not behaviour reading', 'minimal mindreading', and 'full-blown mindreading'. These terms were not used by Bugnyar (2011), but his text makes clear that the first conclusion was concerned with behaviour rules, and the last with full-blown mindreading. My labelling the second conclusion 'minimal mindreading' is more of an imposition but—because eye-object line is a relation between an agent, and object and a location—I think it likely that use of this cue would provide evidence that an animal had, if not all four principles of minimal theory of mind, then at least the first three.

[10] Chimpanzees remain the modal subjects in research on animal mindreading (Whiten, 2013), but in the last 15 years it has become increasingly common to test birds of the corvid family.

[11] Experiment 1 in Bugnyar (2011) also included a condition where one of the competitors saw both caching events and the other saw neither caching event. This 'stay treatment' was not crucial to the interpretation of the results and therefore, for the sake of simplicity and brevity, I have not included it in this summary.

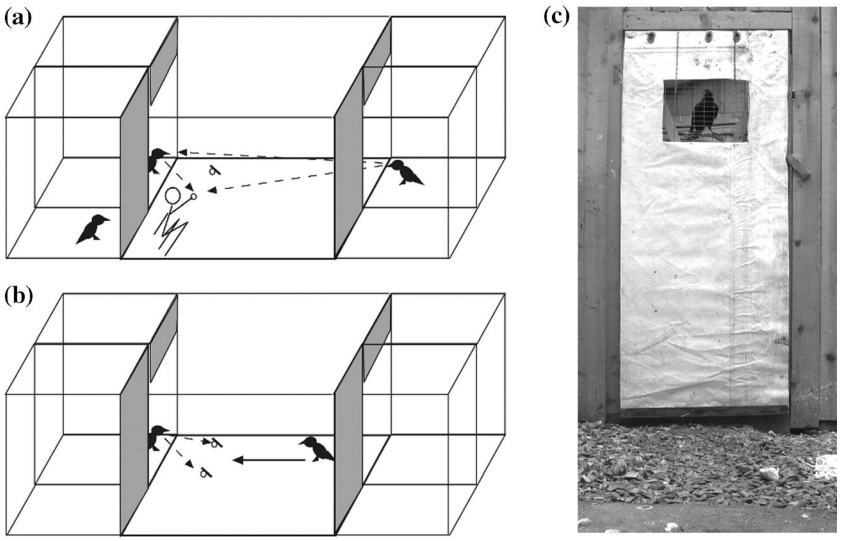

Fig. 3 Experimental set-up used with ravens in Bugnyar's competitive feeding paradigm. Reprinted from Bugnyar (2011)

The first step in Bugnyar's reasoning—from (1) to (2)—illustrates the currently pervasive tendency to regard behaviour rules as the only alternative to mindreading. When one turns to cognitive science and begins to think about alternatives as domain-general, submentalising processes, it becomes clear that the effect of the window manipulation is not sufficient to secure the conclusion that the ravens must have been using minimal (eye-object line) or full-blown mindreading. For example, it is possible that in the retrieval phase of no window trials, the physical appearance of the competitor on the other side of the enclosure (O1 or O2) cued retrieval from memory of the location at which food was cached in the presence of that competitor, and, because approaching the cued location had been rewarded in the past, resulted in the S choosing the cache that matched the competitor. (It is unlikely that the ravens learned within the experiment to approach the cued location because each was given very few trials, but there is no reason why they should not have learned this in the course of their competitive day-to-day lives before the experiment began.) The ravens' failure to match cache with competitor in window trials could have been an encoding specificity effect (Tulving 1983). In window trials, but not in no window trials, the competitor appeared in a different context in the retrieval phase (without a curtain or window) than it had in the caching phase (in a curtain window). Therefore, with fewer of the encoding cues present at retrieval, it is possible that in the window condition the physical features of the competitor bird were less effective in activating a representation of the location of the matching cache.

This submentalising hypothesis could be tested by including additional window trials in which the competitor appears high up in a window, but the location of the cache is such that the lower part of the curtain does not obstruct the line between the competitor the caching event. If the birds are using eye-object line (minimal mindreading), one would expect them to show cache-to-competitor matching in these new window trials, but if the encoding specificity hypothesis (submentalising) is correct, the S ravens should behave in the same way in new as in old window trials.

More broadly, because the submentalising perspective reminds us that the appearance of mindreading can be given by domain- and taxonomically-general processes, it encourages experiments with inanimate targets and unpromising species. For example, in Bugnyar's paradigm, competitor birds O1 and O2 could be replaced with equally salient and discriminable patterns of inanimate stimuli, and the animate and inanimate versions of the test could be given to pigeons as well as ravens. If these tests showed that the cache-to-competitor matching effect is specific to animate targets and/or species that are thought to be specifically adapted for complex social interaction, they would support the idea that this matching effect depends on minimal or full-blown mindreading. On the other hand, if the results indicated domain- and/or species-generality, they would encourage elaboration and testing of alternative submentalising hypotheses.

The second step in Bugnyar's reasoning—from (2) to (3)—was, as he declared himself, speculative. He did not provide, or claim to provide, evidence of full-blown mindreading. However, in the spirit of Bugnyar's move from (2) to (3), I agree that if we had evidence that an animal can use eye-object line as a guide for choice behaviour, it would be well worth testing that animal for something more like full-blown understanding of seeing and/or knowing. But how could this be done? I continue to think that the "goggles" method, which I proposed some years ago, could be helpful in this respect (Heyes 1998; following Novey 1975).

The goggles method could be implemented by converting Bugnyar's version of the competitive feeding paradigm into a procedure involving conditional discrimination training followed by transfer tests: First, the ravens would be given plenty of trials

in which, from an adult human's perspective, what a competitor has seen in the caching phase predicts where a competitor will go when released. To promote generalisation, the cues used in these trials would be physically disparate. For example, they might include old and new window trials, no window trials of the kind used by Bugnyar (2011), and no window trials in which a competitor is seen to have its back turned during caching. Second, in parallel with this training, each raven would be given direct experience with two screens (the analogue of goggles), one transparent and the other opaque. The screens would have salient borders of different colours (e.g. red–transparent, blue–opaque, counterbalanced), but would otherwise not be discriminable when viewed at a distance equal to the width of the central enclosure. The ravens would be given a task in which they look at and attempt to look through these screens, and thereby have the opportunity to discover—on the basis of their own experience only—that the red one affords seeing, and the blue one does not. Finally, the birds would be given trials in the competitive feeding paradigm where, in the caching phase, the S raven could see that the red or the blue screen was interposed between the competitor bird and the caching event. Evidence that ravens can represent seeing could come from a stronger tendency to match caches to competitors in red (transparent) than blue (opaque) trials, when these trials are not differentially reinforced, i.e. when the probability that the S will get the second cache does not vary between red and blue trials. However, a more sensitive test would be provided by training half of the birds with 'seeing-compatible' mappings, in which the probability that the competitor will get the second cache is higher in red than in blue trials, and the other half with 'seeing-incompatible' mappings, in which the probability that the competitor will get the second cache is lower in red than in blue trials. More rapid development of cache-to-competitor matching in the seeing-compatible than the seeing-incompatible group would suggest that the attribution of seeing and/or knowing to competitors promoted learning in the seeing-compatible group, or interfered with learning in the seeing-incompatible group, or both.

Since I outlined the goggles method in 1998, it has been relabelled in a variety of ways (e.g. as the 'opaque visor', 'experience projection' and 'self-other inference' method), applied in research with infants (Meltzoff & Brooks 2008; Senju et al. 2011) and adults (Teufel et al. 2010), and frequently recommended as a nonverbal test of mindreading (e.g. Penn & Povinelli 2007; Whiten 2013), but very seldom used—or at least, reported as having been used—in research with animals. Vonk and Povinelli (2011) reported a pilot experiment in which they tested chimpanzees using a version of the goggles method, and found no evidence of mindreading, but they tested only two or three animals, and gave each only 16–24 trials.

There are a number of potential reasons why so few goggles experiments with animals have been reported. First, it is possible that many have been done but they yielded negative results and therefore have not been published. Second, researchers may have avoided the goggles method because it is technically demanding. For example, one must ensure that the colour cue is the only feature that distinguishes the screens (goggles) when they are viewed at a distance, that the Ss do not see another agent responding to the screens prior to the test trials, and that the colour of the opaque screen does not become aversive to the Ss. It would also be tricky in the competitive feeding paradigm to make sure that Ss could see the physical features of their competitor as well as the border colour of the screen interposed between the competitor and the caching event. There is little incentive to rise to these challenges when less demanding methods are more likely to yield what appear to be positive results, and high impact journals are eager to publish reports of animals with human-like intelligence.

Of course, it is also possible that the goggles method has not been used very often because it is a poor test of mindreading. This seems unlikely given the discussion that immediately followed publication (see Commentary and Author's Response in Heyes 1998), and subsequent endorsements (e.g. Meltzoff & Brooks 2008; Penn & Povinelli 2007; Senju et al. 2011; Teufel et al. 2010; Whiten 2013), but two recent objections to the goggles method are well worth considering, in part because they illustrate the inhibitory role that behaviour rules are playing in research on mindreading (Lurz 2009; Perner 2012).

The goggles method assumes that, in order to pass the test, an animal would have to infer from its own direct interaction with the screens that the red screen affords seeing (it allows an agent on one side of the screen to see objects and events on the other side of the screen), the blue screen does not afford seeing, or both, and subsequently to apply these generalisations to another agent. For example, to think: the red screen is between the caching event and agent X. The red screen affords seeing. Therefore, agent X can see the caching event. Lurz (2009) has suggested that animals could solve the goggles problem using generalisations about "direct line of sight" (behaviour reading), rather than seeing (mindreading). According to Lurz, "to judge that a subject, S, has direct line of sight with an object, O, is to judge that there is no opaque barrier (of a certain size) on the straight line between S's open eyes and O" (p. 309).[12] Similarly, Perner (2012) has suggested that animals could solve the goggles problem using behaviour rules of the form "if there is a transparent object (goggles or screen) between his eyes and the target he will behave adaptively towards the target, otherwise not". Thus, both critics suggest that an animal could pass the goggles test

[12] Lurz (2009) suggests that his concept of 'direct line of sight' is equivalent to my concept of 'eye-object line' (Heyes 1998), but this is not correct. An agent has eye-object line when there is "an *unobstructed*, notional straight line between their eyes" and the object (Heyes 1998, p 113, emphasis added). That is, when there are no objects interposed between the agent's eyes and a focal object. In contrast, on Lurz's account, an agent can have direct line of sight when there is an object between the agent and the focal object, as long as the interposed object is transparent.

using the concepts of transparency and opacity that do not subsume or implicate the concept of seeing, and Lurz has explained in some detail what such concepts may be like. For example: "It is quite plausible that the concept of opacity that chimpanzees (as well as other animals) use to distinguish opaque from transparent barriers/media are primitive (i.e. nondefinable), much in the way that colour concepts are generally taken to be. Thus, for example, a chimpanzee's concept of opacity might simply be the concept C* such that if it sees (or seems to see) an object O behind/ within a barrier/medium Y, then, ceteris paribus, it is disposed to believe that Y is not C*, and if it sees (or seems to see) a barrier/medium Y but does not see (or seem to see) object O but nevertheless believes (based upon the contents of its working memory of the environment) that O is behind/ within Y, then, ceteris paribus, it believes that Y is C*" (Lurz 2011, p 37).

I do not doubt the coherence or ingenuity of these proposals. It is certainly possible in principle that chimpanzees and other animals have concepts of transparency and opacity of the sort described by Lurz, and that they could use these concepts to pass the goggles test. However, unlike Lurz, I do not regard these proposals as "plausible", or as a solid basis for rejecting the goggles test, because they are not supported by independent evidence that animals (or humans) can conceive of transparency / opacity in this way, or by an outline of a practicable experimental design that could produce such evidence. Without this kind of support—which would come from cognitive science rather than common sense—the transparency / opacity proposals merely demonstrate that, like all scientific enquiry, research on animal mindreading is subject to the problem of underdetermination of theory by evidence (Stanford, 2013). These proposals would need the support of cognitive science in order to become, not just in principle possibilities, but alternative hypotheses; to provide the means and the motivation to devise an experiment—for example, using screens with different properties—that could distinguish the transparency/opacity hypothesis from the seeing/not seeing hypothesis. Surely, outside the curious world of mindreading, this is how science works—incrementally, by dealing with each theoretically and empirically motivated problem as it comes. Thus, success in a goggles test certainly would not show us once and for all that the tested animals are capable of full-blown mindreading, but I think it would be a step in the right direction.

In summary, using Bugnyar's (2011) exemplary implementation of the competitive feeding paradigm, I have suggested that experiments using inanimate control stimuli and unpromising species could be helpful in distinguishing minimal mindreading from submentalising, and that—although it is technically demanding, and subject to routine underdetermination—the goggles method has mileage as a means of distinguishing minimal from full-blown mindreading.

Conclusion

So, animal mindreading: what's the problem? I have suggested that methodological factors have contributed to a lack of progress in this field, but that the core problem is theoretical—it is no longer clear what the search for animal mindreading is searching for. I believe this problem can be solved using minimal theory of mind as a source of target hypotheses, and, under the banner of submentalising, cognitive science rather than common sense as a source of alternative hypotheses. If we tackle the problem using these resources, and by returning to the demanding but potentially effective methods developed in the 1990s, there is a fighting chance that we will get an empirically sound answer to the question: How, if at all, do animals read minds?

Acknowledgements I am grateful to Ian Apperly, Cameron Buckner, Thomas Bugnyar, Martin Eimer, Nick Shea and an anonymous referee for their thoughtful comments on an earlier draft of this article.

References

Apperly, I. A. (2011). *Mindreaders: The cognitive basis of "theory of mind"*. Hove: Psychology Press.

Bräuer, J., Call, J., & Tomasello, M. (2007). Chimpanzees really know what others can see in a competitive situation. *Animal Cognition, 10*(4), 439–448.

Buckner, C. (2013). The semantic problem (s) with research on animal mindreading. *Mind & Language,* (in press).

Bugnyar, T. (2011). Knower–guesser differentiation in ravens: Others' viewpoints matter. *Proceedings of the Royal Society B: Biological Sciences, 278*(1705), 634–640.

Butterfill, S. A., & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind & Language, 28*(5), 606–637.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences, 12*(5), 187–192.

Campbell, D. T. (1954). Operational delineation of "what is learned" via the transposition experiment. *Psychological Review, 67*(3), 167–174.

Cheney, D., & Seyfarth, R. (1990). Attending to behaviour versus attending to knowledge: Examining monkeys' attribution of mental states. *Animal Behaviour, 40*(4), 742–753.

Cheney, D., Seyfarth, R., & Smuts, B. (1986). Social relationships and social cognition in nonhuman primates. *Science, 234,* 1361–1366.

Clayton, N. S., Dally, J. M., & Emery, N. J. (2007). Social cognition by food-caching corvids. The western scrub-jay as a natural psychologist. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 362*(1480), 507–522.

Goldman, A. I. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford: Oxford University Press.

Hall, G. (1991). *Perceptual and associative learning*. Oxford: Clarendon /Oxford University Press.

Hare, B., Call, J., Agnetta, B., & Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour, 59*(4), 771–785.

Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behaviour, 61*(1), 139–151.

Heyes, C. (1993). Anecdotes, training, trapping and triangulating: Do animals attribute mental states? *Animal Behaviour, 46*(1), 177–188.

Heyes, C. (1998). Theory of mind in nonhuman primates. *Behavioral and Brain Sciences, 21*(1), 101–114.

Heyes, C. (2012). Simple minds: A qualified defence of associative learning. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 367*(1603), 2695–2703.

Heyes, C. (2014a). False belief in infancy: A fresh look. *Developmental Science*. doi:10.1111/desc.12148

Heyes, C. (2014b). Submentalizing: I'm not really reading your mind. *Perspectives on Psychological Science, 9,* 131–143.

Heyes, C., & Frith, C. D. (2014). The cultural evolution of mind reading. *Science, 344,* 1243091. doi:10.1126/science.1243091

Honey, R. C., & Hall, G. (1989). Acquired equivalence and distinctiveness of cues. *Journal of Experimental Psychology: Animal Behavior Processes, 15*(4), 338.

Hurley, S. L., & Nudds, M. (2006). The questions of animal rationality: Theory and evidence. In S. L. Hurley & M. Nudds (Eds.), *Rational animals?* Oxford: Oxford University Press.

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

Lurz, R. (2009). If chimpanzees are mindreaders, could behavioral science tell? Toward a solution of the logical problem. *Philosophical Psychology, 22*(3), 305–328.

Lurz, R. (2011). Belief attribution in animals: On how to move forward conceptually and empirically. *Review of Philosophy and Psychology, 2*(1), 19–59.

Meltzoff, A. N., & Brooks, R. (2008). Self-experience as a mechanism for learning about others: A training study in social cognition. *Developmental Psychology, 44*(5), 1257.

Novey, M. S. (1975). *The development of knowledge of others' ability to see*. [Unpublished doctoral dissertation, Harvard University.].

Papineau, D., & Heyes, C. (2006). Rational or associative? Imitation in Japanese quail. *Rational Animals* (pp. 198–216).

Pearce, J. M. (2008). *Animal learning and cognition: An introduction*. Hove: Psychology 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: Biological Sciences, 362*(1480), 731–744.

Penn, D. C., & Povinelli, D. J. (2013). The comparative delusion: The behaviorstic/mentalistic dichotomy in comparative theory of mind research. In J. Metcalfe & H. S. Terrace (Eds.), *Agency and joint attention* (pp. 62–78). New York: Oxford University Press.

Perner, J. (2010). Who took the cog out of cognitive science? *International Perspectives on Psychological Science, 1,* 241.

Perner, J. (2012). MiniMeta: In search of minimal criteria for metacognition. *Foundations of Metacognition* (pp. 94–116).

Povinelli, D. J. (1987). Monkeys, apes, mirrors and minds: The evolution of self-awareness in primates. *Human Evolution, 2*(6), 493–509.

Povinelli, D. J. (1994). Comparative studies of animal mental state attribution: A reply to Heyes. *Animal Behaviour, 48,* 239–241.

Povinelli, D. J., Nelson, K. E., & Boysen, S. T. (1990). Inferences about guessing and knowing by chimpanzees (*Pan troglodytes*). *Journal of Comparative Psychology, 104*(3), 203.

Povinelli, D. J., Nelson, K. E., & Boysen, S. T. (1992). Comprehension of role reversal in chimpanzees: Evidence of empathy? *Animal Behaviour, 43*(4), 633–640.

Povinelli, D. J., & Vonk, J. (2004). We don't need a microscope to explore the chimpanzee's mind. *Mind & Language, 19*(1), 1–28.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*(04), 515–526.

Qureshi, A. W., Apperly, I. A., & Samson, D. (2010). Executive function is necessary for perspective selection, not level-1 visual perspective calculation: Evidence from a dual-task study of adults. *Cognition, 117*(2), 230–236.

Rizley, R. C., & Rescorla, R. A. (1972). Associations in second-order conditioning and sensory preconditioning. *Journal of Comparative and Physiological Psychology, 81*(1), 1.

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance, 36*(5), 1255.

Santiesteban, I., Catmur, C., Coughlan Hopkins, S., Bird, G., & Heyes, C. (2013). Avatars and arrows: Implicit mentalizing or domain-general processing? *Journal of Experimental Psychology: Human Perception and Performance*. doi:10.1037/a0035175

Schmelz, M., Call, J., & Tomasello, M. (2011). Chimpanzees know that others make inferences. *Proceedings of the National Academy of Sciences of the United States of America, 108*(7), 3077–3079.

Schmelz, M., Call, J., & Tomasello, M. (2013). Chimpanzees predict that a competitor's preference will match their own. *Biology Letters, 9*(1), 20120829.

Senju, A., Southgate, V., Snape, C., Leonard, M., & Csibra, G. (2011). Do 18-month-olds really attribute mental states to others? *Psychological Science, 22*(7), 878–880.

Seyfarth, R. M., & Cheney, D. L. (2012). Animal cognition: Chimpanzee alarm calls depend on what others know. *Current Biology, 22*(2), R51–R52.

Stanford, K. (2013). Underdetermination of scientific theory. In E. N. Zalta (Ed.), *The Stanford encylopedia of philosophy.*

Teufel, C., Alexis, D. M., Clayton, N. S., & Davis, G. (2010). Mental-state attribution drives rapid, reflexive gaze following. *Attention, Perception, & Psychophysics, 72*(3), 695–705.

Tipples, J. (2002). Eye gaze is not unique: Automatic orienting in response to uninformative arrows. *Psychonomic Bulletin & Review, 9*(2), 314–318.

Tipples, J. (2008). Orienting to counterpredictive gaze and arrow cues. *Attention, Perception, & Psychophysics, 70*(1), 77–87.

Tomasello, M., Savage-Rumbaugh, S., & Kruger, A. C. (1993). Imitative learning of actions on objects by children, chimpanzees, and enculturated chimpanzees. *Child Development, 64*(6), 1688–1705.

Topal, J., Miklosi, A., Gacsi, M., Doka, A., Pongracz, P., Kubinyi, E., Csanyi, V. (2009). The dog as a model for understanding human social behavior. *Advances in the Study of Behaviour, 39*, 71–116.

Tulving, E. (1983). *Elements of episodic memory.* Oxford: Clarendon.

Vonk, J., & Povinelli, D. J. (2011). Social and physical reasoning in human-reared chimpanzees: preliminary studies. In J. Roessler, H. Lerman, N. Eilan (Eds.) *Perception, causation and objectivity.* Oxford Scholarship Online. doi:10.1093/acprof:oso/9780199692040.003.0019

Whiten, A. (1994). Grades of mindreading. In C. Lewis & P. Mitchell (Eds.), *Children's early understanding of mind. Origins and development* (pp. 47–70). Hillsdale: Erlbaum.

Whiten, A. (2013). Humans are not alone in computing how others see the world. *Animal Behaviour, 86*(2), 213–221.

Whiten, A., & Byrne, R. W. (1988). Tactical deception in primates. *Behavioral and Brain Sciences, 11*(02), 233–244.

Woodruff, G., & Premack, D. (1979). Intentional communication in the chimpanzee: The development of deception. *Cognition, 7*(4), 333–362.