The University Of Sheffield.

Mindreading in Animals

By

Yifan Mei

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield
Faculty of Arts and Humanities
Department of Philosophy

November 2022

Acknowledgements

I own many thanks to a number of people. Without them, this thesis would not have been done. I would like to express my sincere gratitude to my supervisors, Stephen Laurence, Rosanna Keefe and Gerardo Viera for all their help and advice with this PhD. They have taught me a lot about philosophy of mind and always introduced me to new ideas. I have also benefited tremendously from their comments. I also owe so much to my parents, especially my mom. She gives me a lot of support from the very beginning to the very end of my whole PhD period. I also want to thank my colleagues and friends in the philosophy department. They have built a friendly and encouraging environment in this department. I always enjoyed the time spent with all of them.

Abstract

Mindreading, or Theory of Mind (ToM), is the ability to attribute mental states like knowledge and belief to others. It is the foundational ability helping people to understand, explain, and predict other's behaviour. It had been thought to be a uniquely human ability, separating human beings from other creatures. However, the question of whether animals have a ToM has been debated since it was first bought forward by Premark and Woodruff (1978). In this thesis, I will argue that some animals, at least some nonhuman primates, including chimpanzees, and some birds, namely corvids (also known as the crow family, which includes birds like ravens, crows and rooks) possess a ToM, which enables these animals to attribute mental states like intentions, perceptions, and beliefs to other animals or to humans. I have called this position *the mindreading account* in this thesis. To justify my account, I have compared it with the three most important alternative accounts. The first alternative account is the *behaviour reading theory*. Advocates of this theory, like Penn and Povinelli (2007b), argue that animals are using behaviour rules to understand and predict behaviour instead of ToM. But I argue that the behaviour reading theory itself is ill-defined, that it should not be treated as the default theory to explain animal behaviour, and that at least for the animals noted earlier, ToM provides a better account than the behaviour reading theory. The second alternative account is the *minimal theory of mind account* proposed by Butterfill and Apperly (2013). This account holds that animals can attribute mental states but disagrees with the ToM account on what mental states they can attribute. In particular, this account wants to keep these states as minimally mentalistic as possible. But I argue that this approach of deflating the kinds of mental states that are attributed is problematic. The minimal theory is neither good enough to explain animals' behaviour in real cases, nor simple enough to claim to be a minimal theory. The third alternative account is the *knowledge attribution theory* (KAT) supported by such theorists as Phillips (2021). This theory agrees with the standard mindreading theory that animals can attribute mental states like perception and knowledge but disagrees with the ToM account with respect to belief attribution. I argue that the evidence given in favour of the theory is flawed, and that the continuing positive false belief test results in animals undermines KAT. After arguing why none of these proposed alternatives can replace the mindreading theory account, I turn to focus on the experimental evidence for ToM in non-human primates (especially chimpanzees) and in corvids. In each case, I provide carefully selected evidence to show that the animals in question have ToM. By investigating the experiments in both chimpanzees and corvids, I conclude that

theory of mind is not uniquely human: it is present in at least two further types of animals—nonhuman primates and corvids.

List of Contents

Declaration

I, the author, confirm that the Thesis is my own work. I am aware of the University's Guidance on the Use of Unfair Means (*www.sheffield.ac.uk/ssid/unfair-means*). This work has not been previously been presented for an award at this, or any other, university.

Introduction

Being able to think about what other people think is a very useful ability in everyday life. Whenever you see other's actions, like moving a chair or holding a door, you might wonder why people are doing that. The answers typically concern how others are thinking and what their intentions or beliefs are. This ability to understand others' action in mentalistic terms is variously known as "mindreading", "folk psychology", "mentalising", or "theory of mind" (abbreviated as, "ToM"). In this thesis, I will use all these terms interchangeably. To be clear, I use a term like "mindreading" to refer to the ability of an agent to track any kind of mental states including intentions, perceptions, knowledge, and beliefs. However, a particular mindreading account will refer to a specific explanation which takes an agent to use particular mentalistic terms to understand others' behaviour. Mindreading is a foundation stone of people's psychological life and social interaction. It helps us to understand, to reason, and to predict others' actions. Like many other human cognitive skills, people wonder whether other animals share this ability with us. And, if so, are they adopting the same theory of mind as us when they understand others' behaviour in mentalistic terms? Answering this question will lead us to understand mindreading in much more detail in both humans and animals. If some animals actually have a Theory of Mind, it will dramatically change our perspective on their behaviours and how we should interact with them. It will also shed light on the evolutionary roots of humans' mindreading ability. It could be the case that our mindreading ability has come a long way from other species. If it turns out that no animals except for humans possess such an ability, it will push us to think where mindreading comes from. For example, is the full capacity for mindreading developed during our childhood, or is it the case human mindreading has a shared evolutionary root with that of chimpanzees? All of these interesting inquiries are informed by how we understand animals' mindreading.

In philosophy, there is a long tradition of scepticism about whether we can really know that other agents have minds, which is often referred to as "the problem of other minds". Of course, I'm not going to call into question whether others really have minds. This is something that researchers studying the question of ToM in humans and other animals already take for granted. But the problem of other minds also raises the question of how we can know what others' minds are like. The most reliable way to know how others think is by directly asking them. On the other hand, our

own ToM can help us with this task without relying on language. But this leads to lots of technical questions about how to tell whether an individual is actually doing mindreading instead of using some alternative method for understanding others' behaviour such as relying on rules to predict behaviour that only appeal to observable elements of the environment and overt behaviours (a method often labelled "behaviour reading"). Such a concern is even more urgent in animal studies. In this thesis, by trying to answer the question whether any animal can mindread, I want to explore other options for understanding animals' behaviour. The most promising cases for supporting mindreading in animals come from the studies of non-human primates (like chimpanzees and monkeys) and corvids (songbirds in the Corvidae family, which include ravens, crows, jackdaws). They are also the most researched animals in this area. So, I will mainly focus on them in this thesis. To answer such questions, I will combine studies from psychology, cognitive science, and philosophy. My strategy is to compare different theories based on current data. I will present three alternative theories to compare to the mindreading account. They are the behaviour reading theory, the minimal theory of mind account, and the knowledge attribution theory. I will argue that none of these theories is a successful alternative to the mindreading account and so conclude that animals – at least chimpanzees and corvids – can do mindreading. Before I talk about animals, I want to briefly introduce some related studies involving human infants. The study of infant and animal ToM share many experimental methods, and they face the same queries from different alternative theories.

Mindreading in infants

There is little doubt that normal human adults possess the ability of mindreading. But the stage at which humans acquire such an ability is still a matter of debate. One standard method to tell whether one agent is capable of mindreading is the false belief task. To succeed on a false belief test, the participant needs to understand that the agent involved holds a false belief about a current situation. The ability to ascribe a false belief to others is a sufficient threshold to say someone can do mindreading. To hold a belief that contradicts the actual situation manifests the idea that beliefs can be different from and even inconsistent with reality. So, the verbal false belief task (the agent can explicitly express such ascription) is widely regarded as the golden standard of testing who possesses mindreading.

Wimmer and Perner (1983) introduced the first version of a false belief task, and later Baron-Cohen et al. (1985) modified it to the Sally-Anne version of the false belief test, which is now a standard verbal false belief test. In the Sally-Anne test, participants watch a puppet scenario and answer question about puppet's beliefs. In the scenario, Sally places a marble into a basket. After she leaves the scene, Anne moves the marble from the basket to the box. Then, Sally returns to the scene and experimenters ask the participants the question "Where will Sally look for her marble?" The correct answer is "in the basket" which requires the participants to attribute a false belief to Sally. The consensus on such tests is that typically children above 4 years old can consistently pass the verbal false belief test. Children around 3 years old consistently fail such tasks (For a systematic review see Wellman et al., 2001). Such evidence has motivated the claim that children under 4 years old cannot mindread others, or at least that they do not have the full capacity of mindreading (Gopnik, 1993; Wimmer & Weichbold, 1994).

However, not all researchers agreed with this conclusion. Some (Ozonoff & McEvoy, 1994; Russell et al., 1991) have argued that other cognitive skills like linguistic abilities and working memory are required to pass such a test. A new group of tasks involving non-verbal false belief tests were introduced around the millennium (Call & Tomasello, 1999). In these new tests, instead of directly asking how children think about others' beliefs, alternative behavioural cues like their line of gaze are considered as indictors. In such non-verbal tests, the setup scenario is simpler than the classic Sally-Anne test in order to reduce the load in working memory. Normally, only one puppet was introduced. In such scenario, an actor watched a toy being hidden in one of the two boxes, and the toy was moved while the actor was not looking. In the testing stage, instead of verbally asking the question, violation of expectation (VOE) and anticipatory looking (AL) are the main alternative measures. Since both these measures are also widely used across animal studies, I will explain them in more detail.

The VOE paradigm is based on the fact that individuals will look longer at when surprising events happen, when their expectation is violated. In the practise, the participants watch either an expected or unexpected event (that is, events that *adult viewers* would take to be expected or unexpected). By measuring the time that children spend looking at these events, researchers can determine the

children have the same sorts of expectations as adults by seeing if they look longer at the unexpected event relative to the expected event. In such tests, if infants can attribute false beliefs to others, they should look longer when the actor reaches for the location where the object actually is because if they actor has a false belief, that isn't where they would be expected to reach. That was what Onishi and Baillargeon (2005) found when they ran this type of test with 15-month-old infants.

The AL paradigm is also based on participants' expectations. The idea is that participants should look first at the place where they expect a relevant or interesting event is going to happen next. Again, in the test, if infants are able to do mindreading, they should anticipate an actor with a false belief will go to the box where she placed her toy, even though the infant (but not the actor) knows that it has subsequently been moved. In test phrase, by tracking infants' eyes, Southgate et al. (2007) found 25-month-old infants did in fact look to the box where the actor had placed the toy, showing that they expected the actor to search in this location based on the actor's false belief.

Since the introduction of these two methods, many different versions of non-verbal false belief experiments employing these methods have been published (for a review, Barone et al., 2019). Overall, the results show that the infants' patterns of attention suggest that infants can track others' false beliefs. However, since the results of non-verbal false belief tests rely on nonverbal behavioural indicators, there is a dispute about how we should interpret these results (which will be discussed later in the thesis).

Mindreading studies in animals

The paradigm of non-verbal tasks opens the possibility for studies of animal mindreading since animals lack the ability to communicate with us in language. In recent years, a wide range of studies have used nonverbal tasks to try to demonstrate that non-human primates and corvids have a basic ability of mindreading,[1] and so are able to ascribe mental states like intentions (Behne et al., 2005), perceptions (Bugnyar, 2011; Drayton & Santos, 2018), even false beliefs (Krupenye et

[1] For a review in chimpanzees (Call & Tomasello, 2008) and animals in general (Krupenye & Call, 2019)

al., 2016) to others. On the other hand, opponents have argued that these results can be explained by other theories that do not take animals to be capable of mindreading, just as some theorists argue that the results in nonverbal tasks with human infants can be explained without taking them to be capable of mindreading. Before I say more about the similarity between the infant and animal studies, I will provide a brief history of mindreading studies in animals.

Research on mindreading in animals started with chimpanzees. One of the most influential studies was by Premack and Woodruff (1978). Their paper was the first to ask the question "does the chimpanzee have a theory of mind?". Their original study was about whether chimpanzees can understand human goals. Their paper, and many subsequent papers by others (Savage-Rumbaugh et al., 1978; Povinelli et al., 1996) found negative results. Chimpanzees showed no sign of any understanding of human's thoughts such as intentions and perceptions. Meanwhile, a similar question was introduced: whether primates were able to ascribe beliefs to other chimpanzees. In these experiments, the tasks used were very similar to the non-verbal false belief tasks used in infants. Most of the studies (Woodruff & Premack, 1979; Povinelli et al., 1990) at that time found that primates failed these tasks, just as they failed the tasks involving attributing mental states to humans. At that time, the studies in chimpanzees followed the classic Sally-Anne false belief test I have described above. They relied on the idea that chimpanzees need to follow human gaze as a cue for finding the food. Such setting of experiments is unnatural because chimpanzees in the wild compete almost only with groupmates for food resources (Wrangham, 1980; Hauser & Wrangham, 1987). But Hare et al. (2000) introduced a new protocol which created a more natural environment for chimpanzees. In this new protocol, chimpanzees were put in a setup which is closer to their natural environment and in a context of competition for food. Two chimpanzees, a subordinate and a dominant, competed over two pieces of food between them. In one situation, one piece of the food was placed in a way so that only the subordinate could see it. The other piece was visible to both the dominant and the subordinate chimpanzee. The results showed that the subordinate chimpanzee was more likely to try to get the food which was not in the view of the dominant chimpanzee. By employing experimental designs that used setups close to chimpanzees' natural living environments (for example, involving competing for food where knowledge of others mental states could give an advantage), many more positive results were found among different species of non-human primates including monkeys and chimpanzees (Hare et al., 2001; Flombaum

et al., 2005). A group of recent studies (Krupenye et al., 2016; Buttelmann et al., 2017; Hayashi et al., 2020) even managed to show that chimpanzees can ascribe false beliefs to others. But we cannot rush to the conclusion that chimpanzees have a theory of mind. Just like the studies in infants, because the studies involve using nonverbal behavioural criteria instead of directly asking participants what they actually think, you cannot simply claim they are mindreading. Alternative explanations that do not involve mindreading need to be ruled out.

When researchers were making progress on non-human primates in mindreading studies, another surprising family of species also joined the discussion. Primates are always the most natural candidates for comparison with humans for many human cognitive abilities because they are the closest living species to human beings in terms of evolutionary relatedness. Also, they are morphologically similar to us. So, it should not be surprising that the main focus of animal mindreading studies has been on chimpanzees and monkeys. But when the corvid family was first introduced into this area by Emery and Clayton (2001), it inspired a lot of discussion. The corvids, better known as the crow family, are the family of songbirds including ravens, jays, crows, etc. They have often been thought to be highly intelligent animals because of their relatively large brains (Jerison, 1973; Emery et al., 2007). Corvids also show some superb cognitive skills compared to many mammals. For example, corvids used tools to get otherwise inaccessible food (Hunt, 1996), even manufacturing novel tools (Hunt, 2000; Weir et al., 2002). When you compare corvids to chimpanzees, it can be a surprise to find they share some important similarities. Both have a complex social life and are smart enough to do many types of complex cognitive tasks. Since mindreading is likely linked to a complex social life and requires a capacity to engage in complex cognitive tasks, it is reasonable to assume that a mindreading ability might be present in corvids as well, and such ability could be very helpful in their social lives.

Due to the obvious physical and lifestyle differences between corvids and primates, the traditional false belief test has to be modified to be carried out on corvids. Most research has focused on the caching behaviour among corvids. Caching food for later consumption is a common practise in birds including many corvid species like ravens and crows. Researchers found that corvids employ highly flexible strategies to protect their caches from pilferers both in the wild and in experimental circumstances (Emery & Clayton, 2001; Emery et al., 2004; Bugnyar & Heinrich, 2005; Dally et

al., 2006). For example, one of such protection strategies is caching the food further away from bystanders who might observe where the food is cached. One particularly interesting question is whether corvids would come up with different ways to protect their caches based on different situations. When Corvids were provided with the opportunity to re-cache their food, they preferred to relocate the food that had originally been cached while they were being observed by a bystander (Thom & Clayton, 2013). When corvids have to cache food in the presence of a conspecific, they preferred to do it in a dimmer location (Dally et al., 2004), or out the view of the conspecific (Dally et al., 2005). Moreover, experiments with jays showed that what they fed their partners depended on what their partners had been eaten earlier (Ostojić et al., 2014; Ostojić et al., 2016). If their partners had earlier fed on one type of food, they would feed them with a different type of food (as jays preferred a varied diet). Crucially, the choice of food for their partner can be different from their own current food preference. Such results suggest the possibility that jays feed their partner based on an understanding of their partner's food desires. This kind of behaviour could be treated as desire attribution. I will present and discuss these experiments and arguments in more detail in later chapters.

The Structure of the Thesis

In my thesis, there are five chapters apart from introduction and conclusion. In the first three chapters, I will discuss the three most influential alternative theories to the theory that animals (and infants) have the mindreading ability. I will argue that the mindreading account is better than all three of these alternatives. Then, in chapters 4 and 5, I will examine the two most promising kinds of animals that might be taken to have the mindreading ability—chimpanzees and corvids—to explore what kinds of mindreading ability they have.

In chapter one, I will present my argument against one of the most influential challenges to the mindreading account from the behaviour reading theory. According to the behaviour reading theory, animals are only reacting to behavioural cues based on behavioural rules. Of all the alternative accounts to the mindreading theory, the behaviour reading theory has the fewest commitments to animals possessing an understanding of any mentalistic concepts. Advocates of such a theory claim that animals are not attributing mental states of any kind to others and that

animals' behaviour can be completely explained in terms of behaviour rules. I will argue that this theory faces three kinds of serious challenge. The first is that the behaviour reading theory itself is a mess. It lacks the simplicity that its advocates assume that it has. The second is that the theory is not self-evident as many assume. Simply recognizing that it is a possible alternative is not enough to show that we should accept it as the right view. Evidence needs to be provided to support it. The third point is that the experimental evidence does not in fact support the theory.

In chapter two, I will present my case against another alternative theory—the minimal theory of mind. Unlike the behavioural reading theory, this theory does take some animals to have the capacity to ascribe some internal states to others. However, these internal states are different from the mental states that we attribute to other human beings using our mindreading ability. The minimal theory involves two alternative concepts, encountering and registration, which it takes animals to use instead of concepts for the traditional mental states of perception and belief. The theory introduces several principles to show how such concepts can explain behaviour from animals and infants without taking them to attribute more complex propositional attitudes to others. I will argue that their claim that infants and animals attributions of mental states to others can be fully understood in these minimal terms cannot be sustained. Their alternative concepts and principles struggle with the current evidence. To fix such shortcomings, they have to expand both their minimal concepts and principles. By doing so, the theory largely loses its promise of being a minimal theory.

In the third chapter, I will discuss the third main alternative to the standard mindreading account. This alternative involves a different version of mindreading: the knowledge attribution theory (KAT). This alternative theory, unlike the previous two, shares quite a lot of ideas with the standard mindreading theory. Both theories agree that animals can attribute perception and knowledge to others. The main difference focuses on whether animals can attribute beliefs to others. KAT argues that since animals have a lot of trouble passing the false belief test, it is better to understand them as not being capable of belief attribution. Their main point is that knowledge or factive states are a more fundamental type of state compared to belief states, which are a non-factive type of state. I will refer the standard mindreading theory as a belief attribution theory (BAT). In the case of this third alternative to the mindreading account, I'm not arguing against the whole theory, instead, I

want to emphasise that KAT is weak version of BAT. But I will argue that if people are willing to admit KAT, BAT is a better choice. Because KAT needs to not only provide a theory to explain how knowledge without belief works, but also provide unique evidence that only supports KAT not BAT. However, they face challenges on both grounds. Plus, the recent experimental evidence in relation to false belief tests in chimpanzees is on the side of BAT.

In chapter four, I will present my case in favour of mindreading in chimpanzees and assess which kinds of mental states they are capable of attributing to others. By introducing standards for what kind of experimental evidence could be counted as good evidence in the beginning, I argue that chimpanzees pass tests that are sufficient to show that they are capable of attributing numerous kinds of mental states to others, including goals and intentions, perception and knowledge states, and beliefs, including false beliefs. The best available theory for all the evidence out there is the mindreading theory.

In the last chapter, I present my case in favour of mindreading in corvids and assess which kinds of mental states they are capable of attributing to others. The corvid family provides another interesting case study of mindreading in animals not only because it offers a totally different experimental approach, but because it also investigates different perspectives on the mindreading ability. In this chapter, I argue corvids, like chimpanzees, are capable of attributing perceptions and knowledge. However, unlike chimpanzees, currently there isn't any investigation on false-belief attribution in corvids. Instead, corvids showed another distinguish mental attribution ability. They are capable of attributing desire or preference to others.

Chapter 1 The behaviour reading theory

In this chapter, I want to talk about one of the most influential theories opposed to the mindreading theory in the study of animal Theory of Mind. I will call it the Behaviour Reading Theory (BRT). The core idea is simple. The theory tries to explain how animals can track others' behaviour in terms of behavioural concepts like stimuli, without reference to mental states or psychological processes. Animals doing behaviour reading can notice statistical correlations between certain kinds of events or stimuli and other animals behaving in particular ways. For example, after few repeats of feeding a dog after the sound of a bell, the dog started to salivate when heard the bell. Here dog's salivation is associated with the sound of the bell. Clearly, certain ideas in the behaviour reading theory are borrowed from behaviourism which has a long history in both psychology and philosophy. However, an important difference from the doctrine of behaviourism which was a general theory of the nature of the mind which aimed to provide objective theorists with a way of explaining behaviour in terms of stimuli and responses (for a review Graham, 2000), the behaviour reading theory only concerns a very limited scope which is mainly in ToM studies, where theorists are only interested in how a given agent explains and predicts the behaviour of another agent in day to day life. But scholars like Penn and Povinelli did expand their arguments further into other field such as causation (Penn & Povinelli, 2007a), and more (Penn et al., 2008). In this chapter, my main arguments are that the behaviour reading theory is problematic at the theoretical, methodological, and evidential levels. In the first section, I will briefly talk about behaviourism in psychology and philosophy. In the second section, I will present the main ideas of the behaviour reading theory. Then, in the third section, I will present the theoretical problems with such an account. In the fourth section, I will argue that the theory is also not supported by the current experimental evidence. Finally, I will provide my conclusion and some suggestions for further studies.

1. A short introduction to behaviourism

Before I present the ideas of the behaviour reading theory, I will briefly introduce behaviourism in philosophy and psychology. Because I think that even if the supporters of the behaviour reading theory don't mention it, some of their key arguments are influenced by behaviourism. By understanding behaviourism, we will be able to have an even clearer picture of the behaviour reading theory.

Behaviourism was an extremely popular research program in 20th century. Philosophers such as Ryle, Carnap, Hempel, and Quine all showed their support for or sympathies with behaviourism. It was even more popular among psychologists such as Pavlov, Skinner, and Watson (Skinner, 1963; Skinner, 1985). The main claims of behaviourism centre around three different types of interests. The first concerns what psychology is; behaviourism defines it as the science of behaviour not a science of mind. While this characterisation is not to deny the existence of the mind — quite the opposite — behaviourists demanded that it was only through the study of behaviour that answers to the question how the mind works could be given. The second concerns the research program inside psychology. Behaviourism claimed that human and animal behaviour should be explained in external physical terms instead of in mental or internal terms. The third concerns how to understand mental terms or concepts. Behaviourists claimed that terms standing for internal states or events should be replaced by behavioural terms or concepts.

Taking one of the most influential behaviourists, B. F. Skinner, as an example, he didn't deny the existence of mental concepts or processes, but he thought all these private processes should be explained in public physical terms by such things as stimuli, responses, learning histories, and reinforcements, and nothing more. For example, a rat would get food by pressing a lever, after several such runs, the rat learned to get the food by manipulating the lever. It looks like the rat knew the fact that in order to get the food it should press the lever. However, according to Skinner, the fact that the rat "[knew] such knowledge" is just a fact about its learning history and nothing more than that.

Behaviourism certainly boosted the development of early psychology research. Learning mechanisms such as classical conditioning and operant conditioning helped to explain many behaviours in both human and animals. But employing only behavioural terms in psychological explanations also limited their explanatory power. There is much behaviour that cannot be explained by the behaviourism approach. More specifically, some cognitive abilities like language cannot be explained in behavioural terms alone. Chomsky (1959, 2013) provided a very strong case that language acquisition cannot be explain by reinforcement learning — that is, in terms of learning mechanisms such as classical or operant conditioning that are central to behaviourism. For example, when children learn to speak a language, the behaviourist would suggest that children learn the utterances by imitating their adults, and their correct utterances are reinforced in the ways like be praised or

get what they ask for. But that certainly is not true. Children can learn to say things that have not been trained to say.

The behaviour reading theory is not behaviourism, as I have noted above. But there is a very subtle relationship between them. In the domain of ToM in animals, the behaviour reading theory essentially takes animals to be thinking like behaviourists. This means that animals can only use behavioural cues to understand other animals' behaviour. For example, Penn et al. (2008) argued that animals were not capable of understanding a higher-level relation based on perception. Animals are trapped at the perceptual level. Such higher-level relation is about the relation between abstract concepts not perception. For example, apples are similar to pears based on their shape. This is a claim about similarity only concerned at perception level. However, when you say apples are similar to pear because of their evolutionary root. This is a higher order claim about similarity. This means that animals can only draw connections between behaviours but cannot appeal to even a single internal state in explaining others' behaviour. They cannot think of others' behaviour in terms of mental states but instead, must rely on only behavioural cues.

2. Behaviour Reading Theories

Since mental states and attributions of them are unobservable, their existence has frequently been questioned in the history of philosophy and psychology. Especially in animals, without the help of languages which are present in the case of humans, it is more difficult to prove that animals possess any particular mental states or cognitive abilities. Theory of Mind is among the most demanding cognitive abilities in animals. Not only does it require the general psychological or cognitive capacity required in to understand and process all of the complexities of animal behaviour, but ToM also requires animals to have highly abstract concepts for mental states to understand how others think. Therefore, many researchers try to avoid attributing ToM to animals to explain their behaviour in experiments. Instead, the behaviour reading account (BRT) explains animals' behaviour by taking animals to explain others behaviour based only on observable cues. In reality, most behaviour reading supporters use the term "behaviour reading" as equivalent to "anti-mindreading". In their arguments, they focus more on arguing *why mindreading is wrong* than on spelling out exactly what the behaviour reading theory is committed to. As Heyes (2015, p. 321) once said "any conditional statement that a researcher can imagine, referring to behaviour and not

to mental states, currently counts as a behavioural rule or strategy." In this section, I will outline several popular theories that might be taken as falling under the heading of "behaviour reading" — accounts that are opposed to mindreading in animals.

2.1 The basic argument of behaviour reading theory

Penn and Povinelli are leading advocates of the behaviour reading theory. One of the most famous papers objecting to mindreading in animals is by them (Penn & Povinelli, 2007b).[2] Penn, Povinelli, and their co-authors provided the most influential and comprehensive argument against ToM in animals. So, I will present their theory as the fundamental theory for any behaviour reading account.

In their 2007b paper, they argued more about why theory of mind accounts should be rejected in relation to animals rather than on explaining how their behaviour reading theory works. Nevertheless, they at least provided a basic model of the behaviour reading theory. They claimed that cognitive creatures already have a basic behavioural learning system which can help them learn from the past experience. We will build to their account in stages.

They begin by providing a very abstract description of any behaviour, b, of a cognitive agent as being a function of several variables, as following:

(1) $b=f(g, r, p, q, \ldots.)$.

This formula basically means that a given agent's behaviour (b) is the product of the interaction of a set of different types of variables. The most significant variables include the following: the internal goal state (g) of the agent, the information states which can affect the goal-direct behaviour (r), the perceptual inputs (p), and the feedback of its sensorimotor loops (q). Variables like g-states and r-states are internal states which "carry information about what the agent has learned about the world that is distinct from the information immediately available to the system's perceptual inputs" (Penn & Povinelli, 2007b, p. 732). This means that for any individual, the information that they can use to guide behaviour does not have to be restricted to just the

[2] Their preference for behaviourism in animal studies did not end with just Theory of Mind. They also express a similar behavioural account of causation cognition (Penn & Povinelli, 2007a). In the paper "Darwin's mistake: Explaining the discontinuity between human and nonhuman minds'' (Penn et al., 2008), they claim that human and animal minds are fundamentally different from each other when it comes to understanding abstract concepts.

information they acquire through perception. The authors argued that the details of this behaviour function are not the key point. The basic idea is that any cognitive behaviour is some function of these variables. Based on this definition of behaviour, the question whether any individual animal possess a ToM is about whether such individual can treat *another* individual's behaviour as the function $b=f(g, r, p, q, \ldots.)$. That is, while the formula given in (1) is a general schema we as theorists have for representing the variables involved in producing behaviour, any animal that is able to engage in mind-reading effectively sees other animals' behaviour in these terms too. So, authors claimed that any mindreader must produce and employ a particular class of information about other's mental states. They called this additional information variable, which is added to the overall description of animals capable of mindreading, *ms*.

According to Penn and Povinelli, this ToM related information should be based on the information already observable to the agent. Using their words "using information observed from the perspective of the system itself:

(2) $ms = f_{\text{ToM}}(r, p, \ldots)$."(Penn & Povinelli, 2007b, p. 733)

That is , *ms* is a special class of information which is a product of the information, *r, p,* etc. Since such information is unobservable, *ms* was inferred from some observable information such as perceptions. In the mindreading theory, individuals need some additional information, *ms*, to predict others' behaviour. Based on Penn and Povinelli's definition in formula (2) and explanation of this definition, I think they would agree that in the case of the behaviour of an agent using ToM, the cognitive behaviour should be defined as follow:

(3) $b = f(g, r, p, q, ms, \ldots)$.

Or they could say the additional information about other's mental states, *ms*, is already in the variable *r* which stands for all the informational states needed. So, at least for mindreading theory, the cognitive behaviour is defined as followed:

(3)' $b = f(g, r(ms), p, q, \ldots)$.

But I think their description here is very misleading and potentially confusing and disguises the difference between the mindreading theory and the behaviour reading theory. The reason is largely due to the authors using the same function to describe two very different things, namely an animal's mind and an animal's attribution of mental states to another animal. The function that describes the elements that interact to produce any cognitive behaviour given in equation (1) should be considered as the general definition of behaviour from the theorist's perspective, not a description of an animal's perspective in observing and explaining the behaviour of another animal. The animal has these states, but unless they have a ToM, they cannot attribute

such states to another animal. But both the mindreading theory and behaviour reading theory are about how to explain others' behaviour from an observer's perspective. The question here is to understand how animals predict or understand other animals. It is not about how we, as theorists, predict or explain animal behaviour. The equation (1) $b=f(g, r, p, q, \ldots.)$ depicts how behaviour should be explained from an external theorist's point of view. So, I think it is misleading to take the mindreading theory to add more variables in this equation, as the equations (3) or (3)' do.

The behaviour supported by ToM, of one animal explaining or predicting the behaviour of another animal in terms of mental states attributed to the observed animal, is a subset of the general behaviour of animals who have theory of mind. It concerns the question how these animals would explain the behaviour of other animals. On the behaviour reading theory, animals cannot explain the behaviour of other animals in these terms. They can only explain others' behaviour in terms of observable variables like *p* and *q*. Here *p* and *q* represent the observable cues such as behaviour cues and external stimuli. While *g* and *r* represent these internal states such as beliefs and intentions. If I put it into an equation, it should like the following:

(4) $pb = f_{\text{behaviourism}}(p, q, \ldots)$,

and the mindreading theory should understand others' behaviour as:

(5) $pb = f_{\text{ToM}}(g, r, p, q, \ldots.)$.

pb here represents "predicted behaviour", and these functions show what variables feed into the prediction of others' behaviour on the two theories. By defining how these theories would predict other's behaviour with (4) and (5) clearly shows the primary difference between them. The behaviour reading should only allow visible information such as the visible and auditory cues. On the other hand, the mindreading theory will not only include the observable information available to behaviour reader but also contain information that is invisible in the sense that it is not publicly observable—such as information goals, intentions, beliefs, and knowledge of others (how much such information should be involved is a matter how much of a full-blown mindreading theory you think is needed).

The second difference is how each theory handles the information available to them. This difference is indicated by the functions: $f_{\text{behaviourism}}$ and f_{ToM}. Each function explains how each theory is implemented in animals. For example, associative learning is a powerful way to explain animal behaviour in terms of stimulus and reward. It could be one mechanism of $f_{\text{behaviourism}}$. For f_{ToM}, many researchers suggest it works in a theory-

like manner. Penn and Povinelli thought that how theory of mind was implemented in animals was not an important concern. But I think that how it works should be taken into consideration.

Penn and Povinelli's account is misleading because the way it is formulated tends to blur the distinction between the general understanding of behaviour and understanding the subset behaviour that involves how animals understand others' behaviour. Penn and Povinelli see animals understanding of others' behaviour in behaviourist terms. But this very different from behaviourism as a general theory of the mind, as I have presented it earlier. Penn and Povinelli admit that internal states play a causal role in animal's behaviour. But they are against the idea that animals are also able to discern the internal states of others. They aren't particularly concerned with how the theory of mind is implemented because they see the key difference between behaviour reading and ToM as that ToM commits more variables into the explanation—those involved in behaviour reading, plus additional ones such as concepts of mental states involved in ToM. But judging whether a theory is the best one should be based on how well it explains the evidence, not just which theory posits the fewest variables. And which theory explains the evidence best will depend on how the theories, including how they are implemented. I will come back to this point in later sections.

To illustrate why thinking of ToM as some additional *ms* is not very helpful, given the differences between these two theories, let us look at their own example (Penn & Povinelli, 2007b). They provided a real-life example to illustrate the difference. The example is as follows: "A chimpanzee (the subject) observes a second chimpanzee turn her head and look off in the distance. In response, the subject turns his head in the same direction." (Penn & Povinelli, 2007b, p. 734)

Penn and Povinelli did not think this example should be count as evidence of f_{ToM} because only using perceptual cues such as eye or face direction is enough to explain such gaze-following behaviour of the subject chimpanzee. There is no need to explain it in terms of inner states such as the belief held by the second chimpanzee that there is something interesting in that direction. In this example, to explain the subject's behaviour, the behaviour reading theory only considered the perceptual cues which were visible from the perspective of the object. But for the mindreading theory, it must rely on some internal states such as other's beliefs. According to their explanation, there is no new variable, like *ms*, needed to explain the chimpanzee's behaviour. However, ToM doesn't ALWAYS have to be used in explaining why an animal reacts to another animal's behaviour the way it does. Some cases, like this one, don't tell us one way or another which theory

is correct. It could be that the chimp follows the other chimps gaze just because following others' gaze has been useful in the past. Or it could be that the ToM explanation is right. This example just doesn't decide between them. And of course, a chimp could have ToM and not use it on some occasions. Just because you have ToM doesn't mean you have to use it all the time. The question here is which function: $f_{\text{behaviourism}}$ or f_{ToM} is better for a given case. And it is not about asking how to justify ToM by identifying *ms* as Penn and Povinelli suggested.

In their paper, they also provided behavioural explanations for the experiments relating to the prediction and explanation of behaviour by corvids (a family of bird species including crows, jays, and others). The behaviour we talk about here is like the cache protective strategies found in corvids such as the caching corvid would remove their caches if another conspecific was watching. They argued that the experimental results from various studies (such as those by Dally et al., 2006, Bugnyar & Heinrich, 2006) can be explained in terms of behaviour rules. According to them, explaining these experimental results in terms of ToM may seem natural to us, but to tell whether corvids actually possess such an ability is not simple at all. In their opinion, the behaviour-reading theory alone is sufficient to explain the protective strategies employed by corvids.

Penn & Povinelli admit that animals like chimpanzees and corvids are capable of many advanced cognitive abilities. They listed three of them (Penn & Povinelli, 2007b, p. 737): 1. "representational architectures of enormous sophistication"; 2. "inferential and simulative mechanisms for forming abstractions about classes of behaviours and environmental conditions that are relevant to their goal-directed actions"; and 3. The ability to "generalise lessons learned from these abstractions to novel scenarios". They think of these abilities together with observable cues suffice to explain the animals' behaviour, while maintaining that animals cannot represent or reason about unobservable cues, such as others' mental states.

Let me give you another perspective on what a behaviour reading account would look like. They should agree that ToM can take unobservable cues (*g*-states, *r*-states) as input, which is captured by the function $f_{\text{ToM}}(g, r, p, q, \ldots)$. The abilities they described are in the function $f_{\text{behaviourism}}(p, q, \ldots)$. I will treat these abilities behind behaviour reading in the term of behaviour rules. By forming such rules, animals can learn from their past experience and apply them in new environment. This is one way to implement the behaviour reading. If write it into equation, it should be like $f_{\text{behaviourism}}=R(p, q, \ldots)$. By taking observable cues like perceptual and

environmental information as input, animals can form different abstract behaviour rules. Claiming animals possess and learn abstract rules makes their theory very different from the traditional behaviourism. For example, corvids will recache their food when a bystander was present during the caching. This behaviour can be explained by a behaviour rule like "recache the food when a conspecific has oriented towards it". Such a rule is only based on visible cues like the orientation of the conspecific. However, to form such a rule, the corvid should make an inference between the loss of their caches and the presence of others. The rule itself is an abstract concept possessed by corvids, but the way corvids learned it is only based on visible cues. In their assumption about ToM, the mindreading theory not only needs all the abilities promised by the behaviour reading but also require some extra stuff like the unobservable information about other mind.

In line with this assumption, the mindreading account requires extra evidence. The right kind of evidence should prove that mental states or any other invisible information states play a role which cannot be explained by the observable cues.

So, the best way to understand what the behaviour reading theory claims is to consider it as a group of behaviour rules. When comparing to the mindreading theory, the main difference between them is on how they enable animals to predict or understand others' behaviour. The behaviour reading theory equips animals with the ability to form behaviour rules and use such rules to interact with others. On the other side, the mindreading theory provides the ability to interpret others in terms of mental states like beliefs. And the basic argument for the behaviour reading theory is that behaviour rules are a well-documented method in animal studies. There is no doubt that animals are capable of such abilities while the mindreading ability has to provide their own evidence.

2.2 Argument of direct line of sight

A more specific argument in favour of behaviour reading is what is called by some people (Okamoto-Barth et al. 2007, Lurz 2009) the "direct line of sight" argument. Others describe the same idea differently. For example, Heyes (1998) called it 'eye-object line'. According to this argument, what the observer represents is the observable relationship between the observed agent and the object. More specifically, the observable cues are (1) that the observed agent is oriented to the object and (2) that there isn't anything blocking the agent's line of sight to the object. Whenever these cues are observed by a bystander, he or she can react appropriately based on a behavioural rule. For example, the behavioural rule: when a raven has a direct line

of sight on the food when you are caching, you should recache it. In principle, this kind of rule could be learned from previous experience. One of the key points in this argument is how to define "direct line of sight" and its relation to potential barriers. Because when an agent has a direct line of sight to an object this means that there are no opaque barriers between the agent and the object along their line of sight. So, what really matters in "direct line of sight" is what counts as an opaque barrier. Normally, the concept of an opaque barriers is based on the concept of seeing. That is, an opaque barrier is one that agents are not able to see through. But a behaviour reading account cannot use this understanding since it requires the theory of mind notion of *seeing*. In attempting to reformulation this rule while avoiding using the concept of seeing, Lurz (2009) provided a very detailed version of the concept of "opaque barriers" that aimed to avoid theory of mind concepts. In his account, the concept of opaque may not be defined by other concepts, but it at least can be functionally defined by its inferential role. According to Lurz, if the subject can see a barrier and cannot see the object behind it, but still believe the object is behind it (based on the working memory of the environment or something similar), the barrier is an opaque one. On the contrary, if the subject can see the barrier and the object, the barrier is not an opaque one. According to memory and inferential ability, chimpanzees can learn what could be counted as opaque barriers like painted walls, wooden doors, and they also know what transparent barriers look like, for example, windows and mesh nets. Of course, since whether a barrier count as opaque or not is based on the experience, animals must have been exposed to the relevant sorts of barriers (opaque and not). So, from their own knowledge of what is an opaque barrier, chimpanzees can judge whether others are in direct line of sight with certain objects. Here, I would accept Lurz's definition of the account for direct line of sight and use his account as the default argument for direct line of sight.

This specific behaviour reading account fits Penn and Povinelli's definition spelled out in the last section. It only takes observable information as input. Information like the direction of other's sight and blockers that are visible from the perspective of the attributer. Take corvids as an example, a direct line of sight behavioural rule could explain the corvids' caching behaviour without reference to any attributions of knowledge states to observers or even states of seeing. For example, in the experiment of Emery and Clayton (2001), the re-cache behaviour of the corvids in private could be explained by direct of sight argument. In the experiment, re-caching behaviour is always related to whether the competitors' have a direct line of their sight to the caches. And it may be supposed that ravens already know the rule that they should recache the food when others have a direct line of sight on the food. Competitors having a direct line of sight is generally a very good indicator

of re-caching behaviour. When the sight is blocked in some experiments (like in the cases of Dally et al., 2005 and Bugnyar, 2011), recaching behaviours disappeared.

What type of experiment might be used to pull apart explanations in terms of the "direct line of sight" and explanations in terms of Theory of Mind? Heyes (1998) proposed a *goggle experiment* that could be done to discriminate the difference between these two theories in primates. This design was also discussed and improved on by Povinelli and Vonk (2003) and Penn and Povinelli (2007b). In the proposal, the chimpanzees would be trained to wear two different types of coloured goggles. From the perspective of the chimpanzees wearing the goggles, the red one is opaque, and the blue is transparent. The only noticeable difference between these two goggles from a bystander's view is the colour of goggles. It is not obvious to bystanders that one type is opaque and the other is transparent. After a chimpanzee is familiarised with goggles as a subject, it is given an opportunity to beg for food from one of the two human experimenters, one wearing the red goggles and the other wearing the blue goggles. If chimpanzees can understand others' perspective, they should beg more frequently from the human who is wearing the blue goggles. The basic logic behind in the goggle experiment is that chimpanzees have no good reason to draw a connection between the colour and the visibility. Because in their experience an object's colour is nothing to do with its transparency. If chimpanzees cannot tell from a bystander point of view that the blue goggles are transparent and the red goggles are opaque, the only way for them to know that the human wearing the blue goggles can see (but the one wearing red goggles cannot) is to project from their own experience. Projecting one's own experience to others should be counted as an inference involving concepts of unobservable. So, the behaviour reading theory cannot explain this phenomenon. The direct line of sight account also falls short of explaining it because the sight was not blocked from the perspective of the attributor. So, the account in the direct line of sight argument does not provide a good reason to selectively beg from one of the humans and not the other, since both of them have a direct line of sight to the subject chimpanzee.

Given the increasing evidence for ToM in corvids in addition to chimpanzees, Heyes (2015) provided an updated version of this type of "goggles experiment" for corvids. In her proposal, windows instead of goggles are the key method to form a brand-new experience in corvids. The basic setup based on an earlier experiment by Bugnyar (2011) (The detail of Bugnyar's experiment is explained in Chapter 5). In the experiment, three ravens (two observers and one focal subject) are in a competition to get the food from two locations stored by a human experimenter. The result shows that the focal raven tended to choose the food which the observer

had been witnessed being hidden (the other food would be safe since the other bird didn't observe it being hidden). When no observers were present when the food was hidden, the focal raven didn't show any preference to the location. Heyes' suggestion for modifying this paradigm is that two different coloured screens are introduced. Just like the chimpanzee case, the screen with red border is opaque and the one with the blue border is transparent. When one of observers and the focal subject confront each other in the central room, they are separated by the coloured screens (a transparent or an opaque one). And the focal raven is given the chance to experience the difference in these two screens and learn that they are able to see through one of them but not the other. If the focal raven is tracking others' perspectives, it should know that the observer behind a blue screen (the transparent one) can see the food location just as it does. So, the focal raven will choose that location to rule out the observer's chance to get it. Correspondingly, when the screen is red (opaque), the focal raven should have no preference to the food location because it knows that the observer cannot see where either cache of food is. On the contrary, if the raven just tracks the direct line of sight of the observers, it should always go for the location which observer has a direct line of sight regardless of which screen it is behind. So, there would be no difference between two screens.

Some scholars like Lurz (2009) are still not convinced by this goggle experiment proposal. He argues that behaviour in the goggles experiment can still be explained in terms of "direct line of sight". For him, the chimpanzees just need one more inference in their mind: "red goggles prevent one from having direct line of sight with objects in the environment, while blue goggles do not" (Lurz, 2009, p. 312). So, in this case, chimpanzees can learn that the red goggles create an opaque barrier during the training stage. Therefore, the begging behaviour in the condition with blue (transparent) goggles is explained by the fact that they have a direct line of sight with the experimenter. Heyes does not agree with Lurz, here the central disagreement is about how to define chimpanzees' learning ability in different situations. For Heyes (2017), chimpanzees are not able to learn that the red goggles create an opaque barrier for them because in their natural environment, chimpanzees don't encounter goggles. And, it is implausible to take chimpanzees have such a flexible learning ability that they could learn this generalization on a single trial with the goggles in the experimental situation. I agree with Heyes. But she did not say much in defence of her argument in her paper. So, I will provide more detail why the goggle experiment would be a counterexample for the direct of line account and behaviour reading in the next paragraph.

If the "direct line of sight" account involves a version of behaviour reading, the positive result from goggles experiment would be argue against the "direct line of sight" account. There is a gap between tracking others' direct line of sight and acting based on some behaviour rules. Even if the mindreader knows whether the agent has a direct line of sight with the object or not, without a specific relevant behavioural rule, there is no reason to suppose that she or he will act appropriately. In the case of birds' caching behaviours, tracking other ravens' direct line of sight does not guarantee that the agent will take protection behaviours. In this case, it seems quite intuitive how the behavioural rule works. Since the past experience of other ravens' having direct line of sight with the caches is statistically correlated with the agent losing these caches, it is better to take some protections. So, the behaviour rule here could be that ravens take certain protection behaviours when other ravens had a direct line of sight with the cache. How the ravens came to have this behavioural rule can be potentially explained in two ways. One of the possibilities is that they are born with it. The other is they learned it from their past experience. Due to the complex situation regarding direct lines of sight in the real world, it is highly implausible for the animals to have behaviour rules of this sort from the beginning. So, animals' ability to track others' direct line of sight must be learned from their own experience. As I mentioned it the last paragraph, the key to direct line of sight is the role of opaque barriers. To estimate whether other agents have a direct line of sight with an object is to estimate whether there is an opaque barrier between them. The key question here turns to how the animals learn the concept of the opaque barrier from their past experience. For those who think the goggle setup is no different from the normal cases, they basically assume that chimpanzees or ravens can learn the red goggle is an opaque barrier from their own experience in a short time. However, goggles are a very different type of barrier to these animals from their familiar ones, since opaque barriers in their native environment always blocks the sight from both first person and third person experience, while the goggles only block the sight from their own experience. If animals can project their own experience to others', then they might think that the other agents' line of sight is blocked by the red goggles even if it is transparent from their point of view as observers. Currently, no evidence shows that either chimpanzees or ravens have such learning abilities. So, I do think the goggle experiment can be used to test the "direct line of sight" account.

3. The problem with the theory

In this section, I will talk about the main problem with the behaviour reading theory itself. I think the theory faces three interlaced questions. First, if we follow the suggestion from Penn and

Povinelli that the details of how behaviour reading works is not a concern in the debate, we will miss a significant part the theory. Second, if we take the behaviour reading theory as claiming that animals have a collection of behavioural rules that guide their interactions with other animals as I presented in previous section, the theory ends up being committed to a lot in comparison to the mindreading theory. The difference in complexity between these two theories turns out to not be as much as Penn and Povinelli claim. Third, if one of the behavioural rules the theory posits is a rule along the lines of the one suggested by the "direct line of sight" account, it will not be enough. The theory will lack the explanatory power promised by its advocates.

3.1 what exactly is the theory of behaviour reading?

Penn and Povinelli (2007b) only gave a general definition of the cognitive behaviour and a suggestion for how the mindreading account should justify their theory. In the previous section, I gave a function about how to predict behaviour based on the behaviour reading account which is the equation (4) $pb = f_{\text{behaviourism}}(p, q, \ldots)$. Also, they claimed the detail of their theory is not important. Such a claim only worked because they thought ToM is the only account where we need to explain how it works. If the general behaviour is understood as the equation (1) which is $b = f(g, r, p, q, \ldots)$, to predict other's behaviours, we do not necessary need more variables like *ms* which is suggested by the mindreading account. So, Penn and Povinelli argue that the burden of providing evidence is on the mindreading account. But as I explained before, both the behaviour reading and the mindreading theory are targeted at a subset of animal behaviour. It is about how animal think about other individuals, how to explain and predict their behaviour, and what action is appropriate in light of it. If the subject animal can understand the behaviour of others in terms of g-states and r-states as equation (1) promised, they are attributing unobservable information to others and so employing ToM. There will be no difference between these two theories. So, the behaviour reading should be presented in the form of equation (4). In this way, two theories can have different explanation of animal behaviour. To answer the question how such behaviour functions can be implemented, I'm not requiring the behaviour reading theory to provide the whole details of such processes. I only ask them to provide a tentative one. Penn and Povinelli underplayed the important of implementation. This leads to a disadvantage for the mindreading theory. Because by highlighting the input to the reasoning process involved in interpreting the behaviour of others posited by each theory, ToM must prove themselves by explaining why they need the extra unobservable variables. It looks like the behaviour reading theory is easier to actualise because animals only need to track observable features.

But in real environment, implementing the behaviour reading theory can be very difficult only based on observable features.

Penn and Povinelli have no problem admitting in their paper that if animals can perform behaviour reading, they should possess some cognitive abilities. Moreover, in another paper, Gallagher and Povinelli (2012) gave a detailed hypothesis regarding how behaviour reading works. They called it the behavioural abstraction hypothesis. It argues that animals should have the following three cognitive abilities: the ability to "(a) construct abstract categories of behavior, (b) predict future behaviors following from past behaviors, and (c) adjust their own behavior accordingly." (Gallagher & Povinelli, 2012, p. 150) This is a simplified version of what Penn and Povinelli listed in their paper (2007b). I think these abilities show how behaviour reading works in more detail. One of the most common ways to implement the behaviour theory is through the behavioural rules. By connecting different categories of behaviour together (or behaviour and environmental stimuli or appropriate responses to animal behaviour), animals can predict behaviour by associating what they learned before. The "direct line of sight" hypothesis is a very specific version of such a hypothesis. If we take the behaviour reading theory as being based on behaviour rules, however, some other problems will come up.

3.2 Behavioural rules and the like

If the behaviour reading theory suggested by Penn and Povinelli is indeed implemented by behavioural rules (BR) or similar hypotheses, the difference between the behaviour reading and mindreading is less than the behaviour reading theory claimed. The key point made by the behaviour reading account was that it meant that agents did not have to deal with troublesome unobservable features when they interact with others. One of the major purported advantages of the behaviour reading theory is that it required less. To decide whether the behaviour reading account is committed to less, the first thing we need to figure out is what they have promised. The fundamental thing is that animals can "form abstract representations of the behaviour of others" (Povinelli & Vonk, 2003, p. 157). For example, "seeing" can be interpreted in behavioural terms as something to the effect that the observed animals have open eyes and their eyes are pointing to a particular place. Since "seeing" is an abstract term, it is not limited to a specific behaviour. On the contrary, when we consider "seeing" as a mental term, it is clear that something extra is needed, like what they can see from their perspective, or from their experience. Since the behaviour reading account cannot refer to mental states, animal behaviour must be explained by a simpler route. This

difference between behaviour reading and mindreading was characterised by Gallagher and Povinelli (2012) as follows.

The inference behind the behaviour reading account (from Gallagher & Povinelli, 2012) is:

Behaviour observations → Behaviour abstraction → Inference to prediction

According to them, ToM adds an extra step beyond those involved in the behaviour reading account, like this:

Behaviour observations → Behaviour abstraction → <Inference to mental state> → Inference to prediction

Their characterisation of the BR account seems reasonable. But I disagree with their formulation of the ToM account and the conclusions they draw from it. In particular, I do not think that behavioural abstraction is needed on the ToM account. Both theories were trying to understand how to get from a set of known behaviours to a set of predicted behaviours. The behaviour reading account used behaviour abstractions and their rules to get this. Instead, the mindreading theory introduces mental states and their references to do the same job. For the mindreading theory, there is not an additional step, instead, it is just a different approach. The mindreading theory is a more demanding theory in the sense that it potentially requires animals to make inferences between mental states and behaviours, while the behaviour reading theory may only need inferences between behavioural abstractions. But even on this point, it is not obvious which account is more demanding. Ascribing internal states to others is more like a ON or OFF question. It is not obvious to me why it is particularly cognitive demanding. The hard part is how to draw the right inference from what you know, in one case about behaviour abstractions, and in the other about the agent's mental states.

3.3 What rules?

If as the behaviour rule theory claimed animals made predictions based on behavioural rules or similar mechanism, there are some questions that haven't been answered.

First, what do these behavioural rules look like? This question concerns how to test the behaviour reading theory. One of the obvious choices is stimulus-response (S-R) rules. Given a stimulus, agents will produce a

particular response. This was illustrated by Pavlov's dog. When the metronome had started ticking, the dog started salivating. This kind of conditioned learning ability alone cannot explain all the behaviour found in animals in experiments and the wild. Because this type of rule is very insensitive to environmental changes. The stimulus, a single behavioural cue, itself is enough to predict the behaviour that followed. Behavioural abstraction is not needed. The connection is drawn from behaviour observation to behaviour prediction, no inference or abstraction is necessary. Besides, rules can be easily controlled in the experimental setup. So, I do not think the behaviour reading theory would solely rely on S-R rules.

Taking the behavioural rule involved in "direct line of sight" as an example, this rule is not defined by the direct line of sight from animals as a single behaviour or even a group of behaviours. There is a behavioural abstraction like direct gazing, but it also considered obstacles. More importantly, how to react to others' direct line of sight is not fixed. An animal could be eating the food under the gazing of others or recaching the food in the same satiation. Penn and Povinelli also provided similar behaviour rules as examples, but they never expressed whether these rules are enough to explain animal behaviour or how many rules are needed. This causes the problem that you cannot design experiments to test the behaviour reading theory because you don't know what to test about.

Second, how are such rules learned? This question concerns the question of how appropriate learned rules are and what their predictive power is. When advocates of the behaviour reading account put forward possible behaviour rules, these rules or relationships are never put to experimental tests. For example, in the case of ravens' recaching behaviour, one of the behavioural rules provided by Povinelli and Penn (Penn & Povinelli, 2007b, p. 736) is "Re-cache food if a competitor has oriented towards it in the past". This rule seems quite intuitive. But whether ravens will actually behave in accordance with the rule is still a question. Is the re-cache behaviour only triggered by the competitor's orientation, or it can be activated by other factors as well? Is the re-caching behaviour the only reaction to the competitor's orientation, or it can take different protection behaviours? As suggested by the current evidence (for more arguments about experimental evidence, please check on Chapter 4), this rule is not enough to cover all the experiment data. Even if the behaviour theory can patch their rules based on new evidence, it signals that the theory lacks predictive power. The behavioural rules are given to deal with a particular result, but they may well not cover other related results and so new rules would be needed for those. But as the same time, the behaviour theory is blinded to what the new rules should be.

Third, a more fundamental problem for the behaviour reading account in relation to animals concerns the structure of the behavioural rules in the theory. Among all the hypotheses favouring behaviour reading, the rules are either too simple or too complex. Hypotheses like the one in the "direct line of sight" account are clearly on the simple side. The hypothesis here relies on a single behavioural rule about the relationship of agents' direct line of sight with the object. But this hypothesis fails to explain several relevant types of experimental evidence (the details of such experiments can be found in section 5 below and more fully in chapter 5). This means it is too simple to explain the findings regarding corvid's behaviours. On the other side, some hypotheses assume that animals generate behaviour rules from their past experience in a very flexible way. It is not about some specific behaviour rules. Instead, it is about the general learning ability in the animals that helps them to act differently in various different situations. People like Penn and Povinelli insisted that this kind of learning ability can also appeal to unobservable features. On their view, animals like birds are able to learn from previous experience and generalise relations between the environment and their goal-directed actions. This type of assumption of a general learning ability in animals has two problems.

The first is that whether birds, for example, have this kind of learning ability is unclear. There has been very little independent experimental work looking into the question of birds' general learning abilities. Beyond the evidence in ToM studies, people do not know to what extent animals are capable of forming rules like connections between behaviours. Because of the lack of understanding of how animals can learn based on situations they observe, the behaviour reading theory lacks the predictive power that the mindreading theory provided. If we grant animals the power of mindreading, they can predict or understand other's behaviour in terms of mental states. Such mental states are shared by themselves and also help to guide their own behaviour. Because animals already behave based on their own beliefs and desires, they do not need extra theories to predict how others would behave if they knew what others believe and desire.

The second problem is the assumption that such learning also involves abstract or unobservable representations just like the mindreading hypothesis does. Penn and Povinelli emphasised that animals possess representational architectures which only involve observable features from their current perspective. However, since animals are able to respond intelligently to novel situations, Penn and Povinelli were forced to hold that animals are capable of forming abstract representations from their past behaviour and that they can make inferences based on such representation. For example, in the case of corvids, ravens can protect their caches

by eating, recaching in different sites, or recaching in a far site. The behaviour theory could use "protective behaviour" to represent such a group of behaviours. The concept of protective behaviour could be considered such an abstract representation of past experience. But by appealing to such a flexible behavioural rule, advocates of the behaviour reading theory lose some of the points that they use to attack the mindreading account. This is because according to them, the undesirable feature of the ToM account was that it relied on inferences between unobservable features. One of the main problems with unobservable is that they are abstractions from the evidence immediately available to an agent and so this evidence underdetermines any rule framed in terms of them. But as I have shown, the behaviour reading theory also adopts certain abstract concepts in their theory. The consequence of this flexible approach is that the behaviour reading just like the mindreading also needs extra evidence to prove that animals do that. By acknowledging a more powerful behaviour rule approach, the behaviour reading theory would face a similar challenge to the one that it raised against the mindreading theory.

If both theories claimed that animals have the ability to represent abstract categories, mindreading also needs observable features to form the representations used in predicting and explaining the behaviour of other animals. The way animals are able to use mindreading is also based on observable features like the direction of others' gaze. The difference here isn't that one of the theories projects beyond the evidence that is immediately available and the other doesn't. The true debate is whether animals track others' perspectives or knowledge states like the mindreading account suggests, or whether they learn highly abstract behavioural rules and draw inferences from them regarding the right action to take in a new environment, as the behaviour reading account assumes. As far as current experimental evidence suggests, ravens are able to take protective actions when a non-partner conspecific is presented visually with no intervening barrier[?] or their presence is inferred from acoustical cues along with some visual cues. Furthermore, the protection should include re-cache, caching in a location that is further away or in dimmer lighting conditions or obscured by an intervening barrier as well. So, the behaviour reading account will need to be at least as complex as (if not more complex than) the mindreading account. It is very unfair to accept the behaviour reading without providing evidence supporting the details of the account, while assuming that we should reject the mindreading theory in the absence of detailed evidence in favour of its account.

4. The methodological problem

Some researchers like Povinelli and Vonk (2003, 2004), Hurley and Nudds (2006), Lurz (2009) think that there is a logical problem with the mindreading theory. The logical problem concerns the methodology in mindreading studies. The problem can be expressed as followed:
For every behaviour that can be explained by the mindreading theory, there must be some observable behaviour on the basis of which the mental state attribution is made, therefore, there is a behaviour reading explanation that can explain that same behaviour.

My response to this so-called logical problem is that it is not a logical problem at all, rather it is a methodological problem. This is an important distinction. Because it will indicate how to solve it. People like Povinelli and Lurz suggest that a new experimental protocol must be devised in which all possible behaviour reading hypotheses are ruled out. But I think no such experimental protocol exists. Surely, a better designed experiment is always appreciated. But the problem in the debate will not go away because of a single experimental protocol. In the following, I will examine what the problem is really about and how it will affect the debate here.

4.1 What is the Problem

This problem has been described as a logical problem is because some researchers thought that this problem is not, in principle, soluble in experiments. For example, Povinelli and Vonk claimed "the problem is not the ingenuity of the experimenters; it is the nature of the experiments." (Povinelli & Vonk, 2003, p. 159) To illustrate such reasoning, it can be expressed in the following steps:

1. An animal, A, predicts that another conspecific B will perform behaviour *q* because B holds a mental state *r*.
2. The reason A thinks that B holds *r* is based on some observational fact *p* which is about B's behaviour or environment.
3. Given this, animal A can just predict that B will do *q* based on *p* alone, without attributing mental state *r* to B.

From the perspective of A, the only observable change in this scenario is the fact *p* and later the behaviour of *q*. This pattern is shown in all experimental evidence for ToM. For example, in the corvid recaching case, the presence of a bystander and the re-caching behaviour of the subject corvid are the main observable facts. No

mental states like "I had been watching during caching food" is out there. More importantly, no such mental states can be observed in any type of experiments. To make this argument work, I think it should add one more premise after 2 and before the conclusion, which would be something like the following: no mental state of B is directly observable by A. This is an important part to rule out mindreading theory. But I do not think this argument works. Because *q* doesn't follow from *p* alone. For the behaviour reading theory, it also needs something like a rule to connect them, which is not presented in the argument. What is presented in the experiment are two independent events: *p* and *q*. Without any additional hypothesis, there is no reason to believe *p* would lead to *q*. The behaviour theory also needs something like a rule to enable the animal A to make a prediction to the effect that *p* is associated with *q*. So, the "logical problem" is not only faced by the mindreading theory but also the behaviour theory. Someone may disagree with this by saying both theories need some rule like or theory-like stuff to bridge independent events, but the mindreading theory still need something more than the behaviour reading theory. But this dispute is about what kind of variables should be included in the bridging rules. What the mindreading theory tries to do is not simply add invisible concepts like mental states based on behavioural rules, instead, it replaces behavioural rules with something else. It is better to describe the difference between the theories as one of path selection.

Some scholars have retreated from such a rigorous description of the problem. They do not claim that the mindreading theory cannot be proven by any experiment. Instead, they hold that the behaviour reading theory should be considered as the default explanation of animal behaviour, and any attempt to support the mindreading theory should first falsify the default theory. For example, Penn and Povinelli (2007b, p. 734) claimed the mindreading theory must "create experimental protocols that provide compelling evidence for the cognitive necessity of an f_{ToM} in addition to and distinct from the cognitive work that could have been performed without such a function." At this point, the argument is much more reasonable compared to the rigorous version of the logic problem, but in this version of the argument, the problem is not a logical one anymore. What is claimed is that animal behaviour should be first be explained by the behaviour reading theory. The mindreading theory can only be justified in light of some specified designed protocols. Such protocols have already been discussed in the section 2. But I don't agree that the behaviour reading should the default option here. The only concerned should be which theory best fits with the evidence.

4.2 It is all about evidence

Even if some researchers have downplayed the logical aspect of the problem, the remaining methodological problem is not as strong as they claimed. Because the behaviour reading theory is not a theory only about merely associating different kind of behaviours. It is not the default answer to any behavioural observation. If there is no default theory here, to be considered as a better theory, both theories should be judged by the overall picture in the available evidence. The point is to find the best theory which can provide a unified explanation for the evidence. Additionally, there are a number of criteria on which theories can be evaluated, such as simplicity, coherence, explanatory and predictive power. So, the problem is also not simply about a certain experimental protocol as suggested by Penn, Povinelli, and others. Maybe a certain protocol can be used to reject a specific behaviour reading or mindreading hypothesis. But it cannot eliminate every version of these hypotheses. Right now, on both sides, the respective hypotheses are still imprecise, which makes them more difficult to rule out. This is especially true for current versions of the behaviour reading hypothesis. As I mentioned in last section, the behaviour reading theory is very vague. So, it is very hard to provide an experimental protocol to defend or refute it. The so-called logical problem is really about the question of how to design an experimental protocol which can distinguish these two hypotheses. The difference between them is also not as the logical problem suggests. The mindreading hypothesis can provide a unified explanation to multiple cases. For example, in birds' caching experiments, when the focal raven's caching behaviour is observed by a competitor, in various different setups, it will take different actions (re-cache the food when the competitor is not here, cache the food in a location further away from the present competitor, or cache the food in a dim place). For the mindreading theory, these various phenomena can be explained in one hypothesis: the raven "knows" that someone is watching them and ascribes to them a perception of their caching, so the raven knows that it needs take some protections or the observing raven will steal the cached food. These protections can take different forms — re-caching the food or caching the food in a more private place. However, for the behaviour reading theory, in every situation, it must provide a different associating rule to explain the behaviour in that situation. For the case of birds, the associative rules will be re-cache the food or cache the food in the further location when a competitor is present. As a result, the behaviour reading account might be seen as less simple and more ad hoc than the mindreading account. So, methodologically speaking, the so-called logical problem is not a problem just for mindreading hypothesis. As long as the behaviour theory is a bunch of associative rules, it will need to be evaluated in holistic terms to determine which theory provides the best overall account of all the experimental evidence, just like the mindreading theory.

4.3 The desired experimental protocol

Even though I don't think a specific experimental protocol can decisively resolve the discussion, I would like to see how the behaviour reading theory would fight back on such evidence. To avoid the objection that their standard of ToM was set too high, Penn and Povinelli provided two experimental protocols which they think would count as proof of mindreading. These two protocols are both based on primate experiments and share the same principle. The first one is about a cooperative task between the primate and a human experimenter. This protocol is very much like one Heyes provided in 1998. I will discuss it later in the following section. The second protocol is about a food competition test between dominant and subordinate chimpanzees. In the test, the basic experimental setup is the same as Hare et al.'s classic protocol (2001) which includes a pair of dominant/subordinate chimpanzees competing for two types of food. To rule out the behaviour reading explanation of the chimpanzees' behaviour, Hare et al. included many situations to test whether the subordinate chimpanzees can continually act as the mindreading hypothesis suggested. Because these situations depended on unfamiliar visible cues, it is very unlikely that the chimpanzees could a have already learned a behavioural role involving such cues. I will talk about a similar experimental design were achieved in the study of corvid in chapter 5.

In this section, I will present 2 types of experimental evidence which would be hard to explain by behaviour reading accounts. The first type is suggested by many supporters from the behaviour reading theory such as Penn, Povinelli and Heyes. The second type is more specific to the "direct line of sight" account.

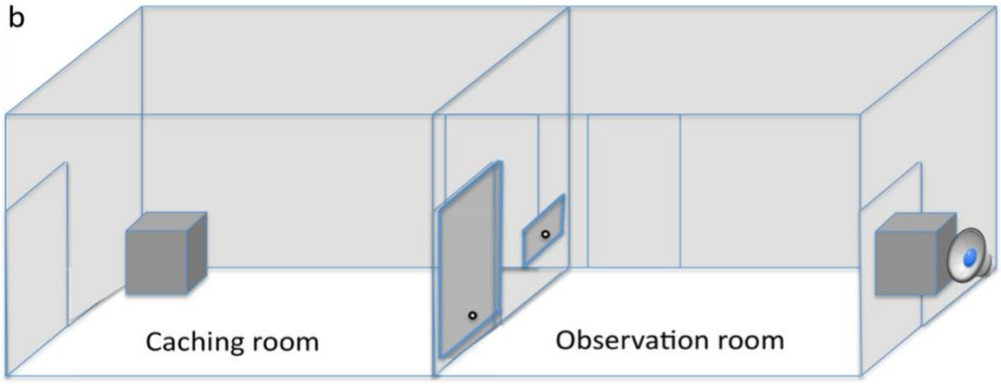
Type 1: The goggles experiment

I presented the concept of goggles experiment in the discussion of the direct line of sight account. In the following I will argue that the experiment by Bugnyar and colleagues (2016) involves a similar test to the test proposed in the goggles experiment required. This type of experiment is a huge threat to both the behaviour reading theory and the direct line of sight account.

In the experiment, two rooms are adjacent and separated by a wooden wall. On the wall, there are two independent windows, and both have peepholes in them. Windows and peepholes can be opened or closed separately. The setup of the room is shown in figure 1. In the setup, windows are opaque barriers which can be opened to give clear sight through them. So, when they are open, the observer and the cacher both have a

direct line of sight to each other . Peepholes provide a very different type of barrier. When a peephole is open, only the observer can see the cacher not the other way around.

Figure 1



Note. This picture was produced by Bugnyar et al. in 2016, showing the experimental rooms' setup. From "Ravens attribute visual access to unseen competitors", by T. Bugnyar, S.A. Reber, and C. Buncker, 2016, *Nature Communications*, 7(1). Supplementary Figure 1 p.1. (https://doi.org/10.1038/ncomms10506)

The experiment works as follows:

Familiarisation with the windows: all ravens were made familiar with the effects of opening and closing the windows. At this point they were not made aware of the existence of peepholes.

After familiarization with the windows, ravens were put into two conditions:

In observed condition, the windows are open, and the raven caches the food in the presence of an observer. In non-observed condition, windows are closed, and the raven caches the food without being seen by any observers. In both situations, ravens in the observation room made sounds which were audible to the cacher.

Then, the human experimenter introduces the peepholes to ravens. Each raven was introduced to use the peephole to watch a human store the food in the next room. Only once the raven successfully recovered the human-made cache, can the next trial proceed.

Peephole condition: ravens were given food to cache in one room while the windows were closed but one of the peepholes was open. Meanwhile, a loudspeaker was playing the sound of competitors recorded from the non-observed condition (so no competitor was in the observation room, but it will have sounded as if one was).

The result of the experiment was that the cacher's behaviour in peephole test is similar to when they were observed by a competitor in observed condition when the window was open. Researchers concluded that "these data provide clear evidence that raven social cognition cannot be reduced to behaviour-reading" (Bugnyar et al., 2016, p. 3). I will argue that this experiment is the first kind of evidence which meets the requirement from Heyes (1998, 2015) and Penn and Povinelli (2007b).

First, let us check whether this experiment is in line with the goggles experiment proposal. As I outlined in the section 2, the goggles experiment adaptation for birds suggested by Heyes includes opaque and transparent windows (marked by different colours). In Bugnyar's experiment, the coloured goggles are replaced by a normal window and a peephole. The key point in goggles experiment is that the goggles provide a unique experience for corvids. The goggles are special in two ways. On the one hand, animals never experience goggles in their natural environment. So, if ravens behave differently based on different goggles, this result should be the result of their own experience with goggles in the experiment. On the other hand, the goggles are special sorts of barriers which offer a different view from first person perspective and third person perspective. The peepholes in Bugnyar et al.' experiment have the same features as the goggles: birds don't have any experience with peepholes and only the birds that have had a first-hand experience with them can know how they function.

Does the Bugnyar et al. experiment provide evidence against the "direct line of sight" account? Yes. If re-caching behaviour is associated with the direct line of sight from an observer, there is no way the focal raven can track the observer's direct line of sight in the peephole condition. Because in the peephole condition, the windows are closed and only the peephole is open. So even if corvids are using visible environmental cues to guide their re-caching behaviour, it cannot be based on clear evidence that there is a direct line of sight. Because in the peephole condition, the peepholes do not allow cachers to tell if others have a direct line of sights of them.

Next, let's consider this experiment in relation to the general behaviour reading theory proposed by Penn and Povinelli. According to them, in the goggles experiment, the chimpanzees must formulate analogies between first- and third-person behaviour. This is beyond the scope of what behaviour reading enables. The same situation is true of the peephole condition. The cachers acted as if they were being watched. They seem to have realised (by whatever cognitive abilities they possessed, including ToM or not) the following:

1. When they observed through the peephole, they can see what is happened in the caching room,
2. This general epistemic information will be experienced by any other conspecific (this is the procedure of analogy),
3. An observer who possessed this information would pilfer the food in the caching room,
4. So, they should take action to avoid the pilferage.

Penn and Povinelli agreed that the analogy procedure is beyond the explanatory power of the behaviour reading account because it involves a relation between two unobservable experiences.

However, there could be a simpler association-based explanation, such as that the cacher may associate the peephole and a sound of a bird with a threat or an observer. When they find the open peephole and the sound, they would then take the protection behaviour. In principle this kind of association can be established based on any behavioural cues, but I would consider such association to be highly unlikely. The association here is the peephole and the sound indicating an observer or the direct line of sight from another bird. But no such association was observed before birds learned how the peephole worked. Birds would not treat the peephole and the sound as a threat in nature. Can birds just learn such association in the familiarisation trials? The problem is they that they never experienced such association in the trials. They did not look through the peephole from the caching room to find other birds in the next room. What they learned should only be that they can look through the peephole from the observation room. The only reasonable way to form such an association is by analogy. If I can look through the peephole, others also can. As long as this analogy between one's own experience and other's experience is drawn, the behaviour reading theory would struggle.

Type 2, evidence against behaviour reading accounts based on "direct line of sight"

Shaw and Clayton (2013) demonstrated that Eurasian jays will limit the auditory information they provide to conspecifics during caching and pilfering. In their experiment, there were two types of caching trays. One had a noisy gravel substrate and the other one had a quiet sand substrate. The jays will hide the food in three conditions: an alone condition in which no observers are present, a seen and heard condition in which observers had visual and acoustical information of the caching, and a heard-only condition in which potential pilferers can hear but cannot see caching. Each condition included two trials: the quiet trial (caches in quiet trays) and the noisy trial (caches in noisy trays). Shaw and Clayton compared jays' caching preference between the two different conditions. And they measured the performance of the pilferers by two standards. The results showed that in all conditions jays tended to place the food in the quiet substrate. While jays reduced

the portion placed in the noisy substrate in the heard-only condition comparing to the other two conditions. Between the "alone" and the "seen and heard" condition, there was not such preference. So, the researchers concluded that jays are able to use acoustic information from the caching as a cache-protection strategy.

It is very clear that corvids in the heard-only condition did not have any visual cues including direct line of sight at the caches. So, there is no way that the "direct line of sight" theory can explain the preference for the quiet substrate. "Direct line of sight" theory is only suitable for the cases including visual cues. Any evidence showing corvids employed observable cues other than the visual ones will be provide counterevidence against this account. On the other hand, behaviour reading theory in general is not necessarily undermined by such evidence. An advocate of the behaviour reading account can easily come up with additional behavioural rules that cover the non-visual circumstances. For example, jays will cache the food in a quiet place when an observer is present.

In sum, Bugnyar and his colleagues (2016) presented some new positive evidence for mindreading in ravens which meets the suggestion of the first protocol from Penn and Povinelli (2007b) and Heyes's goggles proposal (1998). They were agreed such experimental protocol would constitute a serious challenge to the behaviour reading theory. Second, Shaw and Clayton (2013) provided evidence showing that corvids can use acoustic information for the likelihood of pilfering. Which means direct line of sight is not the only observable cues guiding corvids' caching and pilfering competition. I would not claim these two experiments are enough to refute all behaviour reading theories, but they provide strong counterevidence against versions of the behaviour reading theory that I presented in this chapter.

Summary and further studies

The behaviour reading approach to understanding animal behaviour is a very tempting and fruitful approach in many cases. But in this chapter, I argued that the behaviour reading theory is not suited for ToM studies in animals like chimpanzees and corvids. The problem rises from three aspects of the theory. Firstly, the theory is ill defined in the first place, and it has been mistakenly taken by the supporters to be the default theory to explain animal behaviour. Secondly, since we shouldn't take the behaviour reading theory as the default to understand animal behaviour, both theories should be under scrutiny. Advocates of the behaviour reading theory have wrongly claimed that

only advocates of the mindreading theory need to take the responsibility of providing evidence for their account. Based on all these arguments, I conclude that the behaviour reading theory is poorly presented by its supporters, and it seriously lacks the evidence needed to support it.

Chapter 2 The Minimal Theory of Mind Account

The minimal theory of mind account (which I will also call “the minimal theory”, for short) defended by Butterfill and Apperly (2009, 2013) aims to offer a minimal theory which provides adequate explanations for basic mindreading ability, such as comprehending other people’s beliefs including false beliefs. Unlike the behaviour reading theory I discussed in chapter 1, the minimal theory does not deny animals can represent the internal states of other agents. Rather, this theory only aims to provide an account in which animals attribute only minimalist internal states to others in explaining their behaviour, which enables animals to succeed on ToM tasks without attributing the rich mental states involved in standard ToM explanations. The minimal theory is based on the idea that the ability to track perceptions, or knowledge states, or beliefs does not necessarily require a full-fledged ToM. There is the possibility that some other theory can do the trick as well as ToM can, especially in infants and animals who appear to have a limited ToM. In this chapter, I will argue that the minimal theory of mind cannot account for a large range of data where animals might be taken to be engaged in mindreading. The minimal approach underestimates the difficulty of explaining this data on the basis of such a meagre foundation. If it keeps to its minimalist promise, it lacks the flexibility to cover the overwhelming evidence against it. If it is extended as the evidence demand, it cannot remain a simple and coherent theory.

The first and foremost question about the minimal theory is how to interpret the term “minimal”. Butterfill and Apperly’s effort to build a minimal theory can be seen as having two aspects. The first part is about propositional attitudes. According to them, the most troublesome concept in the mindreading theory involves propositional attitudes like beliefs. Propositional attitudes are normally associated with human language. So, they want to remove propositional attitude attributions from the mindreading process. To do this, they take the minimal mindreading ability to employ more "minimal" psychological notions. In particular, they employ notions like “encountering” and “registration”—which we will get to in a moment—instead of richer notions like perception and belief. By defining these notions in their own terms, they aim to avoid difficulties associated with taking less able agents like animals to attribute propositional attitudes.

The second element involved in their building a minimal theory concerns how to design the theory. Based on the literature of ToM, there are mainly two ways to explain the ability of mindreading. One is called Theory-Theory (TT) which holds that mindreading operates by employing a collection of lawlike generalizations (Gopnik & Wellman, 1992; Gopnik & Meltzoff, 1997). The other one is Simulation Theory (ST) which holds that people use their own mental states to simulate others' (Gallese & Goldman, 1998; Goldman, 2006). The strategy deployed by minimal theorist follows the basic idea from Theory-Theory. In both theories, the mindreader utilises rule-like principles to explain the minds of others. The difference is that the minimal theory of mind wants to keep the rules that animals use as minimal as they can. So, they propose only five detailed principles in their theory.

This chapter contains four sections. In the first section, I will introduce the basic structure of the minimal theory, focusing on the basic principles that Butterfill and Apperly provide to as the basis for being a minimal mindreader. In the second section, I present the basic commitments and limitations they propose to keep the theory a minimal one. In the third section, I will demonstrate that their original principles cannot live up to their promises. In the fourth section, I point out that the minimal approach faces further challenges which are not easy to overcome.

1. The minimal theory of mind account

The minimal theory of mind account, like other theories in this field, attempts to explain the cognitive ability of tracking others' mental states, such as their perceptions, epistemic states, and beliefs. For Butterfill and Apperly, infants and even some animals show a basic ability to track other agents' perceptions and beliefs. For a standard (that is, non-minimal) theory of mind, tracking perceptions and beliefs requires many cognitive resources and higher-order cognitive abilities like causal reasoning, while infants and animals likely do not meet these requirements. As a result, Butterfill and Apperly try to construct a theory less dependent on sophisticated psychological resources. They deploy principles which enable attributors to use simple rules instead of mental principles. I will analyse the theory in two respects. First, I examine the crucial simplification that involves replacing attributions of perceptions and beliefs with attributions of more minimal states

— "encounterings" and "registrations". Second, I examine the particular set of principles that unites these notions into a theory.

1.1 Key concepts

Before going through the detailed principles that comprise the minimal theory, let us elaborate the two vital notions in the minimal theory: encountering and registration. The notion of encountering is similar to that of perception. It shares many basic features with perception, though encountering does not involve more controversial features like representation. Encountering just focuses on the common-sense relation between the subject and the object which is constrained by spatial and physical features. To make this notion more accessible, another notion, "field", is introduced by Butterfill and Apperly. The field of an agent consists of objects. Which objects fall into the field are determined by the orientation of the agent and the circumstances of objects like whether there are barriers between the object and the agent. The considerations which limit one's field include but are not confined to the lighting and spatial situation of the objects and the relational situation between the agent and objects. Based on the concept of field, Butterfill and Apperly define encountering as having an object in the field. So, attributing an encountering to another agent involves taking that agent to stand in the encountering relation to the object. For example, when you see your friend reaching for an apple in front of you, you can attribute she is encountering the apple. By representing what others can encounter, individuals can track other's perception without perspective taking (that is, without actually attributing perceptions to others). Let's look at a more complex example now. In the classic case of Hare et al. (2001), food was placed behind a barrier so that only a subordinate chimpanzee could see it. The dominant chimpanzee in the scenario is unable to perceive the food. So, the standard mindreading account can take the subordinate chimpanzee to not represent the dominant chimpanzee as perceiving the food. In this situation, the subordinate chimpanzee will go for the food since it is able to track the perception the dominant chimpanzee. By using the concept of encountering, the subordinate should also do the trick. For the minimal theory, because the food is out the field of the dominant, the dominant cannot encounter the food. By represent what the dominant chimpanzee can and cannot encounter, the subordinate is similarly able to go for the food. By representing what the dominant can encounter, the subordinate effectively is able to track the perceptions of the dominant without representing these perceptions *as* perceptions. This shows that it is possible to track another's perceptions using

only a minimal theory of mind, rather than a full theory of mind that would involve project oneself into the other's perspective and attributing perceptual states to them.

Next, the notion of registration works similarly to beliefs although registration just records the relationship between an agent and certain objects in a certain location. The authors definition of registration is that "an individual registers an object at a location if and only if she most recently encountered it at that location" (Butterfill & Apperly, 2013, p. 617). So, attributing a registration to another just involves representing that an individual stands in a certain sort of relation to an object and location—namely, one of having most recently had that object in one's field at that location. Let's see how this plays out in relation to a typical false-belief task. In the standard false belief task introduced earlier, Sally put her toy in a basket and then left the room. An agent employing the minimal theory would say that Sally registered her toy in the basket, and so has a representation of her toy being in the basket. Apperly and Butterfill argue that based on the real situation, Sally's registration could be a correct or incorrect registration. If the toy stays in the basket until Sally comes back later, Sally will have a correct registration. But if the toy is moved to a new place without Sally forming a new registration of the object in this new location, then Sally held an incorrection registration. By representing what Sally has registered, an observer can track her beliefs. However, registration removes all the intentional elements from belief. In registration, the attributor does not need to assume the agent has any intentional ability like believing.

1.2 Principles

The minimal theory of mind is constructed with four principles. Apperly and Butterfill (2013) offered four principles, but they do not rule out the possibility of additional principles or amendments. I will focus on their four main principles. However, later in the chapter, I will explore various possible extensions and variations on these principles that could be introduced to address objections facing the theory. For the moment however, I will focus on the four principles that, according to them, are sufficient for an agent be able to track others' perceptions, beliefs, and false beliefs.

1.2.1 First Principle

In their own words, Butterfill and Apperly's first principle is that "bodily movements form units which are directed to goals." (Butterfill & Apperly, 2013, p. 614) This principle embodies the first type of assumption that agents with a minimal theory make in minimally understanding others in mental terms. It links bodily movements to "goals", which might be understood as outcomes to which an action is directed. This principle enables the agent to track other's goals without representing and understanding other agents as having intentions. For example, a chimpanzee walks toward a banana and grabs it in their hands. This action can be interpreted as a goal-directed action: the goal of this chimpanzee is to get the banana. Butterfill and Apperly's aim is not to give a fully adequate account of what goal-directed action actually consists of. Instead, they want to point out that goal-direct actions can be understood in a minimal way without actually understanding others as having rich mental states like intentions. Using this principle, an agent can track others' goals by linking bodily movements with outcomes they are directed at.

1.2.2 Second Principle

Butterfill and Apperly's second principle is framed in terms of their theoretical notion of encountering. It holds that "one cannot goal-directedly act on an object unless one has encountered it" (Butterfill & Apperly, 2013, p. 615). In other words, an agent who uses this second principle of the minimal theory will assume that if an agent has not encountered an object, the object cannot be the goal of the agent. Since encountering broadly tracks perception, an agent who uses this principle will effectively withhold from attribute a goal to act on an object to other agents when the other agents can't perceive that object. And again, they can do this without employing the richer notions in a full (nonminimal) mindreading account. Notice also that this principle can help the agent to prevent other to achieve their goals. For example, Dally et al. (2004) found that ravens prefer to cache food in locations which are out of the view of competitors. This can be explained by principle two by noting that the ravens are trying to avoid having others encounter their food, since according to principle two, others can't act in a goal directed way toward the food without first encountering it. This principle can also help agents to achieve their own goals. For example, recall that in Hare et al. (2001) I have presented in chapter 1, when the subordinate and the dominant chimpanzee competed for the food between them, they found that subordinate

chimpanzees went for the food out sight of the dominant chimpanzee. This can be explained by assuming, in accordance with principle two, that the subordinate chimpanzees understand that this food can't be the goal of the dominant chimpanzee because that chimpanzee hasn't encountered it.

1.2.3 Third Principle

Butterfill & Apperly's third principle is that "correct registration is a condition of successful action." (Butterfill & Apperly, 2013, p. 617) This principle like the second one, can be used both to help an agent achieve their goals and to impede other agents from interfering with their goals. An agent can infer what the agent must have registered from the fact that the agent successfully achieves a goal. Meanwhile, the agent can also predict an agent's action toward an object when observing that the agent has registered the object in the correct location. Unlike the second principle which can only be applied in the situation other agent didn't encounter the object, the third principle also works in the condition where the other agent has encountered the object. For example, in the same experiment of Hare et al. (2001), if the subordinate chimpanzee sees that a dominant chimpanzee has registered where food has been hidden, the subordinate will not go for the hidden food. The second principle had no use here, but the third principle can explain it. Subordinates can realise that the dominant had a correct registration of the food, so the dominant can successfully get the food. In this case, it is better not to confront the dominant competitor.

1.2.4 Fourth Principle

Butterfill & Apperly present their fourth principle as followed: "when an agent performs a goal-directed action with a goal that specifies a particular object, the agent will act as if the object were in the location she registers it in" (Butterfill & Apperly, 2013, p. 619). This principle is specifically designed to address situations like those involved in the false-belief tasks. Thinking about the Sally-Anne task, at the beginning, Sally puts the toy in the basket. After she has left, Anne moves the toy from the basket to the box. Where should Sally look for the toy? Based on fourth principle, Sally should look for the toy in the location where she last registered it as being, which was the basket. In the classical mindreading theory explanation, the participant should ascribe a false belief to Sally. The minimal theory explains it instead by ascribing an incorrect registration to Sally. Another important supplementary statement for the fourth principle is that the agent only needs to

maintain a record of the most recent registration an agent has. In the Sally-Anne tasks, Sally can be thought of as simple registering the toy as being in the basket no matter how many times she has seen the toy moved before her final registration of the toy before leaving. The agent using the minimal theory of mind does not need to keep a history list of all the previous registrations that Sally might have of the toy, since Sally can only act based on the last registration.

1.2.5 Extensions and Variations

Apperly and Butterfill do not rule out the possibility that more principles might be added to the minimal theory or that there might be modification to their original four principles. But they do think these four principles are good enough to cover most situations including the false-belief tasks. Regarding possible extensions to these principles, they list some potential examples. For instance, they suggest that registration might be extended to other types of properties of the object. Moreover, additional principles could be added to support more abilities analogous to those in classical ToM, for example tracking desires. Regarding variations of the current principles, they give an example relating to corvids' protection behaviours. Unlike chimpanzees, corvids caching strategies are not about preventing others from pilfering a particular food item, but instead are about preventing others from pilfering any items at all. They argued that corvids may tracking relations between locations and food types rather than particular items.

In sum, on the minimal theory, when agents show abilities to track others' perceptions and beliefs in tasks, they do not need to be taken to be representing others' perception and belief as the mindreading theory suggests. Instead, they can be understood as representing others' encounterings and registrations. Consider a false belief task in chimpanzees (Krupenye et al., 2016) as an example, involving a setup like that involved in the human version false-belief task. In the false belief scenario, a human agent tries to retrieve an object in a cage, and an ape-like character prevents the retrieval by hiding the object in one of two locations. Under the observation of the human agent, the ape-like character puts the object into one location. After the agent has left the room, the ape-like character moves the object to the other location. The question is where would subject chimpanzees anticipate that the human agent will look for the object. The results showed that chimpanzees will anticipate the human agent acts as though they have a false belief, which means the agent looks for the object in the original location. Based on the principles I have

presented above, the minimal theory can also come up an explanation. The chimpanzee subject can predict the human action by representing what the agent has encountered and registered. First, the chimpanzee understands the goal of the human agent is to get the object in the cage (based on the familiarisation trials that the chimpanzee observes). In those familiarisation trials, the human agent consistently reaches to the object. Then, in the test, chimpanzees understand that the agent will register the object at the original location not the location the object was moved to because the agent last encountered the object in the original location. Last, chimpanzees will predict that the agent will act based his most recent registration. This means the agent will look for the object in the original location where he last registered it as being. This prediction from the minimal theory is the same prediction that the full mindreading theory would make about this false-belief scenario.

In the next section, I want to discuss how the minimal theory can be seen as an intermediate theory between the behaviour reading and the mindreading theory, and what advocates of this theory need to commit to in order to hold the theory to its minimalist promise, and what kind of limitations this implies.

2. Limitations of the minimal theory

Apperly and Butterfill make it clear that the minimal theory of mind is different from both the mindreading theory and the behaviour reading theory. They highlight some limitations regarding their theory to demonstrate this. To make it clear, I want to examine how the minimal theory is different from the other two options and figure out what the minimal theory needs to be committed based on its core principles. Meanwhile, these commitments of the minimal approach will also establish the limitations of this theory in comparison to a full (nonminimal) mindreading theory. The primary limitation emphasised by Apperly and Butterfill is about identification. But I will argue that there are further limitations in their theory beyond those that they acknowledge. To fully understand the limitations of the minimal theory, we need to compare and contrast the theory with other theories, namely the behaviour reading theory and the mindreading theory.

2.1 Comparing the minimal theory to the behaviour reading theory

From the perspective of the behaviour reading theory, Apperly and Butterfill provide several possible ways to distinguish these two theories. According to them (Butterfill & Apperly, 2013), one way to distinguish them, which could be used with human subjects, is to design tasks that involve directly asking the subject to explain how they think instead of just observing their behaviours. For example, in a standard Sally-Anne false belief test, experimenters can just ask the subjects why Sally searched as she did. If the subject humans respond with terms like "Sally thinks" or similar ones, this at least shows that people use a theory of mind to explain other's behaviour. The problem with such a strategy is that it only works where the subjects can answer the 'why' question. Since many subjects that we are interested in, like infants and animals, certainly cannot do that, they need to have alternative other methods. One such alternative method asks whether the subject is able to differentiate their own perceptions and beliefs from others'. Apperly and Butterfill think the behaviour theory cannot accommodate this. Consider the challenge posed by false-belief tasks. The subject holds a true belief about where the toy is, while the other individual holds a false belief about the location of the toy. The ability to distinguish these two beliefs is the key to passing such false-belief tasks. A related test is the "self-other inference", which means that the subject can use their own experience to infer how others would behave in a similar situation. One example of this type is provided by Emery and Clayton (Emery & Clayton, 2001). They found that only corvids who had had the experience of pilfering will take actions to protect their items. When birds without such experiences of pilfering cache food in front of a bystander, they won't take actions like re-caching the food to protect the items (while those who have had such experiences of pilfering themselves, will re-cache the food to protect it). Results like these can be explained in terms of a "self-other inference". When storers had the experience of stealing food from other's caches, they will use such experience to infer that others can do the same to them. This will explain the experimental results from Emery and Clayton.

These ways of attempting to distinguish the behaviour reading from the minimal theory face two problems. The first problem is that I don't think that a "self-other inference" test can really separate the minimal theory from the behaviour reading theory. The second problem is that the way they frame the difference is ambiguous, and can be understood in two very different ways. On the first point, the behaviour reading theory can explain the result from Emery and Clayton (Emery &

Clayton, 2001) without appealing to a self-other inference. For example, they can appeal to an association of the presence of a bystander with the loss of caches to explain the result. If the caching bird hasn't had the experience of pilfering, it will fail to associate these two things together. On the second point, how we interpret the principles offered by the minimal theory will hugely affect how it works. If we take a pro-mindreading approach, all such distinguishing about self and other is achieved by representing encountering and registering. If such representation is considered to be about the internal states of others, the difference between these two theories is obvious. As I have noted in the previous chapter, one of the most important claims from the behaviour reading theory is that agents can only use external cues like behavioural ones to predict others' behaviour. But from a pro-behaviour perspective, encountering and registering might be understood as something like behavioural rules. On this understanding, the four principles in the minimal theory can be treated as behavioural rules. Taking the second principle as an example, "one cannot goal-directedly act on an object unless one has encountered it". If encountering is essentially seen as only involving a more elaborated version of the direct line of sight, the principle is not much different from a behavioural rule like: an agent cannot goal-directly act on an object without that object having been in its direct line of the sight. Also, as I have argued in the previous chapter, the so-called self-other inference can also be accommodated. The behaviour reading theory can come up with certain rules to explain the difference between self and others, at least in the case of corvids.

These sorts of ambiguities in the minimal theory are largely due to the authors' reluctance to commit to behavioural explanation involving attributions of traditional psychological capacities like belief and perception, while trying to retain many of benefits of appealing to them. The authors want two core concepts, encountering and registration, to work like perception and belief. But by their definition of these two concepts is closer to behavioural concepts. On the other hand, their willingness to commit to more by being open to extensions and variations regarding these principles shows that they actually think encountering and registration may involve much more than what is encompass in their rather sparse definitions. This ambiguity serves their purpose in allowing them to claim that the theory works in a minimal sense. But in reality, as we'll see, the only way to make the theory work will be to bring it much closer to a mindreading theory through extensions and amendments.

2.2 Comparing to the mindreading theory

On the other side, Butterfill and Apperly also claim that the minimal theory is different from the mindreading theory. It defines what the agent can represent in terms of encounterings and registrations of other agents. By characterising encountering and registering in a minimal sense, they hope to significantly reduce the psychological demands on attributions of such states as compared to attributions of mental states on the traditional approach which is based on propositional attitude attributions. I want to talk in some detail about these limitations and will divide my discussion into three aspects: one each focusing on encountering, registration, and identity.

2.2.1 Limitations on encountering

Apperly and Butterfill emphasise that "the encounterings are relations not representations" and only representing perception—not representing encountering—"involves representing representations" (Butterfill & Apperly, 2013, p. 616). According to their theory, encountering is about the relation between the agent and an object in its field. The most obvious limitation for encountering then is that the agent cannot represent *the way* in which another agent encounters the object in his or her field. Being able to represent the way in which another agent represents things is also known as level-2 perspective taking. This type of perception tracking is about how the agent can understand how others see the object from their own viewpoint. For example, a flower can be seen in the left side of a tree as viewed from one perspective or the right side of the tree as viewed from a perspective on the opposite side of the tree. This spatial relation between the flower and the tree depends on the viewpoint. If a mindreader sees a flower as being on the left side of the tree, another agent standing opposite to him or her will see it differently. When the mindreader can adopt the viewpoint from the agent on other opposite side instead of their own viewpoint, she or he is doing level-2 perspective taking. In this situation, when the mindreader can only represent encountering, s/he can only represent the flower is also on the left side of the agent.

There are also other limitations on encountering besides the absence of level-2 perspective taking. The concept of encountering is heavily related to visual information. It lacks the ability to cover other forms of perception for example involving auditory and olfactory information. To enable the

agent to track perception beyond vision, the minimal theory would need to expand the concept of encountering. One way to do this is by introducing more field-like concepts. For example, an agent can encounter the sound when an acoustic source is in the sound field of an agent. By computing whether the agent is in the sound field of the acoustic source, the attributer can compute whether the agent encounters the sound or not. But the notion of a sound-field would need to be spelled out fully in terms accessible to agents without mindreading abilities. The situation can be more complex when the judgement involved are based on a combination of different types of sensory information. For example, ravens may use a combination acoustic and visual information to judge whether their caches were exposed to a bystander.

2.2.2 Limitations on registration

One feature of registration is that an agent employing a minimal theory of mind only can register the most recent encountering another agent has had with an object. To reduce the cognitive demands on registration, Apperly and Butterfill (2013) suggested that to register the most recent encounter, an agent using a minimal theory could simply forget all the previous encounters but only retain a representation of the most recent one. Doing this will hugely reduce the demands on the agent's memory capacity. But if will also dramatically restrict the information an agent has about what other agents have registered. Other possible limitations on registration concern the kinds of properties that could be included in a registration. The original definition of the registration from Apperly and Butterfill only mentions that an agent can represent the relationship between some other agent and the target object and its location. No other properties of the object apart from its location are included in the registration.

2.2.3 Limitations on identity

The primary limitation differentiating the minimal theory from classical mindreading that is highlighted by Butterfill and Apperly concerns identity. According to them, beliefs, as propositional attitudes, can track not just that others represent objects, but also *how* others represent objects, while registration can only track the relations between agents and objects. Furthermore, they claimed that this difference between the theories will be manifest in situations

that involve false beliefs involving identity. I will talk about this limitation in much more detail in the later section.

3. Problems concerning the original principles

I will begin by considering the original four principles provided by Apperly and Butterfill (2013) and whether the minimal theory given in these terms is good enough to explain animal behaviour. Because they are claimed to be enough to explaining what is happening in false-belief tasks, without any extension or variations. Being able to do that is itself is a very high bar for any theory trying to explain animal behaviour in the field of mindreading studies. But I will show later in this section, that the minimal theory falls short on its original promises. The problem is based on their key concepts: encountering and registering. These two concepts are defined in a way that allows them to mimic the function of perception and belief. But because they promise to provide on minimal mentalistic factors, these two concepts cause the minimal theory to fall short. I will divide my discussion of problems for the minimal theory understood in terms of the original principles into problems centring around encountering, and problems centring around registration.

3.1 Problems centring around encountering

To understand the problems associated with the notion of encountering, I look at encountering in the context of the four principles, because the function of encountering is only elucidated through these principles. Without the help of these principles, encountering is unable to explain any behavioural expectations animals might have. According to the notion of encountering, when there is an opaque barrier between an agent and an object, the attributor will represent that the agent cannot encounter the object because the object is not in the field of the agent. To clarify a bit, to represent someone as not encountering something can be understood in two ways. On the one hand, it could mean that the mindreader represent the agent has not encountered the object. It is a representation that asserts a negative content. On the other hand, it could mean that the mindreader doesn't represent the agent as having encountered the object. This would involve a lack of positive representation. Here I'm talking about the first situation. Moreover, considering what is a necessary condition for the possibility of seeing an action as a goal-directed action, if the agent has not encountered a given object, the agent will not take any goal-directed actions toward that

object. However, I will introduce two new experiments involving corvids and chimpanzees to show that the concept of encountering is not enough to explain animal behaviours given this understanding of the principle. In both cases, animals must represent individuals as being able to perform goal-direct actions without representing them as having encountered the target.

3.1.1 Experiments about scrub-jays/ ravens

Butterfill and Apperly argue that scrub-jay caching behaviour from experiments (Dally et al., 2004; Clayton, 2007) can be explained by the minimal theory. For example, when observed by a competitor, scrub-jays prefer to cache the food further away from the observer, or in a location which cannot be seen by the competitor. Instead of explaining this caching preference in terms of representing what others can see, the minimal theory explains it with encountering. Because the jays can represent what others encountered, they "know" whether the competitor has encountered the food or not. They also "know" that other competitor can only target their food based on having encountered the food. So, for the caching jays, to protect their food from pilfering, they need to minimise the chance of others' encountering the food. The behaviours found in the experiments, such as caching the food further away from the competitor, are one way to reduce the chance of others encountering the food. One of the troubles that the behaviour reading theory has in trying to explain experiments like these is that for every protection strategy found in corvids, the behaviour theory needs to come up with a new rule. The advantage of the minimal theory is that they can explain different strategies observed from corvids in protection behaviour by appealing to the concept of encountering.

However, in a recent caching experiment, Bugnyar, Reber, & Buckner (2016) show that ravens can learn to track unseen competitors. I have presented the details of this experiment in the previous chapter. You can find more detail there. The basic question of this experiment is under what situations the ravens would re-cache their food items.

We have already known that when raven cache food in the presence of competitors, they tend to re-cache the food later. This phenomenon could be explained by the minimal theory. The caching ravens can represent that the competitors have encountered the food items during the caching stage. This encountering makes the competitors capable of goal-directed actions directed toward the food

items, namely stealing the caches. To prevent the competitors from stealing the food, one thing the ravens caching them can do is re-cache the items. Following principle two of the minimal theory, if the competitors haven't encountered the food, they have no chance to pilfer it. Therefore, the caching ravens don't need to re-cache the food in these circumstances. But in the experiment, they designed a peephole situation. In this situation, the cacher had no visual access to the competitor, but the competitor can see where the cacher hides the food from a peephole. Compared to a similar non-observed condition, where there is no competitor presented during caching, only a noise was played next door to the caching room. The difference between these two conditions was that the peephole was only opened in the peephole condition not in the non-observed condition. After the experimenters showed the ravens how to use the peephole, the cacher ravens will recache their food in the peephole situation but no in the non-observed condition. This result showed that the caching ravens in the peephole condition behaved in a way similar to the situation which they have been observed by competitors.

In both the peephole and non-observed conditions, however, no competitors were in view of the caching ravens. According to the definition of encountering, the cacher cannot encounter the competitor. It is reasonable to think without encountering the competitor, the cacher cannot represent the encountering of a nonvisible competitor. The reason behind this inference is that if the agent can represent the perspective of a nonvisible agent, why introduce the encountering concept at first place? The core benefit of encountering is that the attributer can easily figure out what others can encounter from on his or her own field. The concept of field makes sure that both the attributer and the attributee shared the same field. Based on this notion of a field, the attributer can easily figure out what the attributee has encountered in this field without representing their perceptions. Going back to the experiment, because the competitor is out the view of the cacher, it cannot be taken to have encountered the food, and so no goal-directed action (here this means pilfering) can be performed. This means that there is no need to take protection strategies like re-caching to prevent such pilfering. The result from the non-observed condition supported such reasoning, namely no encountering, no recaching. But the peephole condition showed something surprising for the minimal theory. Even without the competitor having encountered the caches from the perspective of the cacher, the cacher still acted as if the competitor would try to pilfer the

food. At least, based on this finding, the original principle explicating the concept of encountering fails.

However, if we make some small amendments to cover this type of case, principle two of the minimal theory can still survive. One such amendment that could be adopted by minimal theorists would involve enhancing the notion of encountering. This would involve a modification of both the notion of encountering and the related principles. Let's assume encountering records not only the visual information but also the acoustical information. One way to do it is to have the attributor take both certain kinds of sounds in the vicinity of the target object and the presence of an agent as implying that an agent has encountered an object. Therefore, in the raven's case, the cacher represents the sounds as entailing that the observer has encountered the food. So, in the peephole condition, the cacher will act as if there is a visible observer. However, this won't completely solve the problem. In the non-observed condition, the storer can still hear the sounds, but acts like it is not under observation — that is, it does not suppose that a competitor is likely to pilfer the food. To patch this shortfall, the minimal theory needs to come up with a solution which enable the attributer to associate the open peephole and the sound together as the sign of a competitor has encountered the food it has cached. To achieve that, the minimal theory needs to introduce a new concept to cover sound and new principles to cover how auditory and visual information can be integrated together. Even if they are able to do this, that would not be the end of the story though, since there are other sensory channels beyond vision and audition. So, more amendments to the notion of encountering and the associated principles are needed. The extension to original principles is do-able, but the extent to which it can still claim to be a minimal theory is unclear.

3.1.2 Experiments with chimpanzees

Butterfill and Apperly also discuss an experiment studying about chimpanzees that we have already discussed. Recall that Hare, Call and Tomasello (2001) found that the subordinate chimpanzees tend to approach food when they know that the dominant chimpanzees have not seen the food being hidden. Butterfill and Apperly believe that this could also be explained by the minimal theory. On this account, the subordinate chimpanzee can be seen as representing that the dominant chimpanzee has not encountered the hidden food items. Since encountering the food items is a necessary condition of achieving the goal of obtaining these food items, the subordinate

chimpanzee can expect that the dominant chimpanzee will not have the goal of obtaining this food. As a result, the subordinate chimpanzee prefers to obtain the food that was hidden where there was an opaque barrier prohibiting the dominant chimpanzee from encountering it, rather than attempting to get the food which is in plain view of both chimpanzees and so will have been encountered by the dominant chimpanzee.

However, other experiments investigating chimpanzees' mindreading ability tell a different story. Schmelz, Call and Tomasello (2011) argue that chimpanzees can predict competitors' behaviours without seeing their choice. In the experiment, chimpanzee A needs to choose from two opaque boards on a table for a piece of food. One board is flat and the other one is oblique. Without the presence of the conspecific B, the chimpanzee A prefers the oblique board (since the chimpanzees know that the tilt in the board is caused by hidden food). In a contrast condition, a conspecific B appears opposite of A and can choose one of the two boards before A is allowed to choose between them. When the competitor B is choosing, chimpanzee A cannot see the choosing process of B because A's sight is blocked by an opaque plate. According to the minimal theory of mind, in the contrast condition, the chimpanzee A cannot see the competitor B's choice, so A should act as in the previous situation, which means that A should choose the oblique board again. However, the result shows that, in this condition, A does not have a preference for either board. A's action seems to suggest that it can infer the competitors' choice even if they haven't seen them making the choice. To explain this result, A needs to represent B's action toward the board without encountering its actual behaviours. So from the experiment, chimpanzees can predict others' goal-directed actions choosing the oblique board without representing others' encountering. This case is even more difficult for the minimal theory to explain it. Unlike the raven's case, in the contrasted situation, the attributor (chimpanzee A) can neither represent the encountering of chimpanzee B nor represent any other clues, such as sounds.

Above all, these experiments illustrate that both ravens and chimpanzees still can represent others' goal-directed actions toward an object even without their encountering of the object or observing others encountering the object. These attributors have the highly limited evidence to use in trying to understand other agents' minds. However, they still can track other conspecific's goals which

are related to certain objects, even though they have evidence of these conspecifics encountering the objects that would allow the minimal theory to explain their behaviour.

3.2 Problems centring around registration

For the minimal theory of mind, the attributor is able to track the beliefs of an agent through representing their registrations. Unlike belief, registration just records the relationship between the agent and the location of certain objects. Therefore, when the attributor tries to track other agents' beliefs, s/he represents their registrations instead of beliefs. There are two ways to apply the principles of correct and incorrect registration. In one direction, the attributor can predict other agents' actions towards to an object based on representing agents' registrations of the object. Moreover, if the registration is correct, the agents can achieve the object as their goal; otherwise, agents will fail to get the object by acting on their registration. In the other direction, the attributor can infer others' registration about the location of the object through their successful goal-directed actions.

In the following paragraphs, I will analyse the shortcomings of the minimal theory connected to the notion of registration from two directions. One of the challenges is how to track beliefs about objects beyond beliefs about their location. The other one concerns the ability to track identical-looking objects using registration in false beliefs. Some of the experiments I discuss in the following have only been done with the human infants. But since the minimal theory treats infants and certain animals as the same in their theory this work with infants is highly relevant to evaluating their theory overall.

3.2.1 Beyond location

The first problem is how to track some fundamental properties of the object other than locations in a goal-directed action. The content of beliefs is not restricted to information about the location of objects. It often concerns other features, for example, the noises and the colours of objects. Additionally, emotions of the agent also can be crucial to understanding a goal-directed action. Many recent experiments show that preverbal infants can track this information to predict the behaviours of the agent exhibiting these emotions. These will be a problem for the minimal theory

of mind because the theory suggests that the mechanism deploying encountering and registration is enough to explain these basic goal-directed behaviours.

(1) Tracking sounds

Olineck and Poulin-Dubois (2005) find that 14- and 18-month-old infants are more likely to regard a sequence of manipulations about some toys following the sound “There” as an intentional action. In the experiment, the exhibitor will do same similar manipulation with toy devices. In one trial the exhibitor saying "Whoops" shows that this manipulation is accidental. In another trial, exhibitors saying “There” suggests that the operation is intentional. In the test trial, infants have the chance to handle these toys before them. The result shows that the infant prefers to imitate the manipulation after watching the trial with the sound of "There" rather than the one with the sound “Whoops”. If the infants are not able track the sounds made by exhibitors, they will represent registrations of the exhibitor as the same in both trials. Because according to the principles related to registration, the infant would only register the relation between the exhibitor and the location of the toy she interacted with. So, according to such principles in the minimal theory of mind, if the infant identified the manipulation of the exhibitor as her goal, they should just copy such manipulation in both “Whoops” and “There” conditions. While the result is contradicted with minimal theory’s prediction. For this experiment, to represent the registration referring to object and location is not enough to comprehend the exhibitor's behaviour. In order to explain the result, the infants at least need to represent the registration of the sounds made by the agent in her action towards an object.

(2) Tracking colours

Luo and Beck (2010) illustrate that 16-month-old infants can represent a presenter’s colour preference. In familiarisation trial 1, there is a red toy pepper on the right (according to the observer's view) and a black cup on the left. Then the presenter uses her finger to point to the red toy pepper and looks at it simultaneously. In familiarisation trial 2, the setup is the same except changing the toy pepper and cup with a red pyramid and a yellow toy house. Later in the test trails, the presenter either points (looks as well) at a red square on the left or a green square on the right. The result shows that infants feel more surprised when the presenter looks and points at the green square on the right side. This means that the infants predict the presenter will point at the red object

ignoring its location. A reasonable explanation for these results is that the infant can represent other's colour preference in this situation. If we try to use minimal theory to explain these results, we are confronted with difficulties. According to this theory's understanding of goal-directed actions, the infant can treat presenter's behaviours as a successful goal-directed action and represent that the presenter registers the goal of presenter as the object on the right side. So, the minimal theorist can have two predictions in the test stage. One is to suggest that the presenter will points to the right side since she has registered the location previously. The other prediction is that the presenter will choose the objects by chance because in light of the change of object the infant will not have any expectations regarding the goals of the presenter. Neither of these predictions matches the result of the experiment. This experiment suggests that infants can track presenters' beliefs about the colours of objects as well as their location to attribute preferences to them and track the goal of the presenters.

(3) Tracking facial expressions

In one of the experiments conducted by Egyed, Kiraly and Gergely (2013), they demonstrate that 18-month-old infants can predict the protagonist's preference depending on her facial expression. In familiarisation trials, two different objects are placed before the protagonist one on each side. Then the protagonist shows a joyful facial expression when she is looking at one of the objects and displays an abhorrent face when she turns to the other object. When the infant is asked by the protagonist to give one of the objects in the test phase, the infants choose the one related to positive facial expression significantly more often. In this test, infants should know that the aim is to pick out the object preferred by the protagonist. For the standard theory of mind, infants can comprehend the preference of the protagonist through her facial expressions and her relationship with the object. As for minimal theorists, in order to predict the action of protagonist related with her goal, the infant should represent her registration. The registration of the protagonist would only concern the location of the object, then the infants should just attribute the protagonist registered both objects. Since the protagonist's choice has nothing to do with the location of the object, the minimal theory would not be able to make a prediction about how the protagonist will behave. The issue for the minimal theory of mind is that the location information between the protagonist and the object alone cannot represent her goal.

In accordance with these three experiments above, the infants can track more information than the location of an object. For the minimal theory of mind, if it wants to keep the principles of encountering and registration, these principles will at least need to be extended to be able to allow for tracking of things like an invisible agent, the sound made by the agent, the emotions of an agent and the colour of an object. However, supplementing the minimal theory to account for all of these factors is not enough. There are more and more new experiments that showing the variety and flexible of mindreading abilities in infants and animals. The real pivotal problem is the minimal theory of mind's assumption that infants and animals have only a very limited mindreading ability. If (as these experiments argue) infants can track beliefs and preferences about things other than objects' locations, it is hard to believe that they can do this just by employing mechanisms of the sort suggested by minimal theorists.

How might advocates of the minimal theory respond to these three issues? One way to attempt to respond to them is to make the notion of registration, and the corresponding principles, more complex. Minimal theorists could argue that, registrations can register not only the location of an object but also the sounds made by the agent, the colour of the object. and the facial expressions of the agent. The goal-directed actions of an agent whose registrations are represented in these richer ways might then be tracked in multiple ways. There would also be further questions concerning how this kind of registration might work and what additional principles might be needed to govern registrations involving these different factors, and how they would predict an agent's actions toward an object in concert with tracking the goals of the agent. As aforementioned experiments showed, in representing one's registration, sometimes the noise of the agent is most important to understand agent's action, at other times the colour of the object plays a central role. How would the attributor know, under different situations, which factor is the most important one when registrations involve so many different features about agents and objects? The minimal theory of mind would need more principles to explain the how these factors work with each other in registration. When the theory is supplemented to address all these issues, it seems that it will be very hard to tell the difference between registrations and beliefs.

Another strategy an advocate of the minimal theory might try is to build more principles involving categories other than encounterings and registrations to explain these experiments. For example,

in the case of tracking colours, minimal theorists can add a "note" principle. So, the attributor represents that the agent registers the location of the object and “notes” the colour of it as well, where noting is somehow explicated in minimal terms analogous to those used to explicate encountering and registration. The defect of this strategy is that minimal theory will likely require many new principles due to the fact that there are more and more new discoveries it will need to explain. In the end, it will be hard to discern the difference the way in which principles need to be multiplied on the minimal theory to accommodate these results and the way that behavioural rules need to be multiplied to explain them in the behaviour reading theory.

4. Problems with the limitations taken to be central to the minimal theory

As I noted in the previous section, Apperly and Butterfill outlined two limitations characteristic of their minimal theory in comparison to the full ToM account. These two limitations were in relation tracking other’s level-2 perspective taking and in relation to tracking the identity of the object in a false belief test. In this section, I will critically examine each of these limitations. I will argue that though there are only a limited number of studies bearing on them, the existing evidence does not favour the minimal theory.

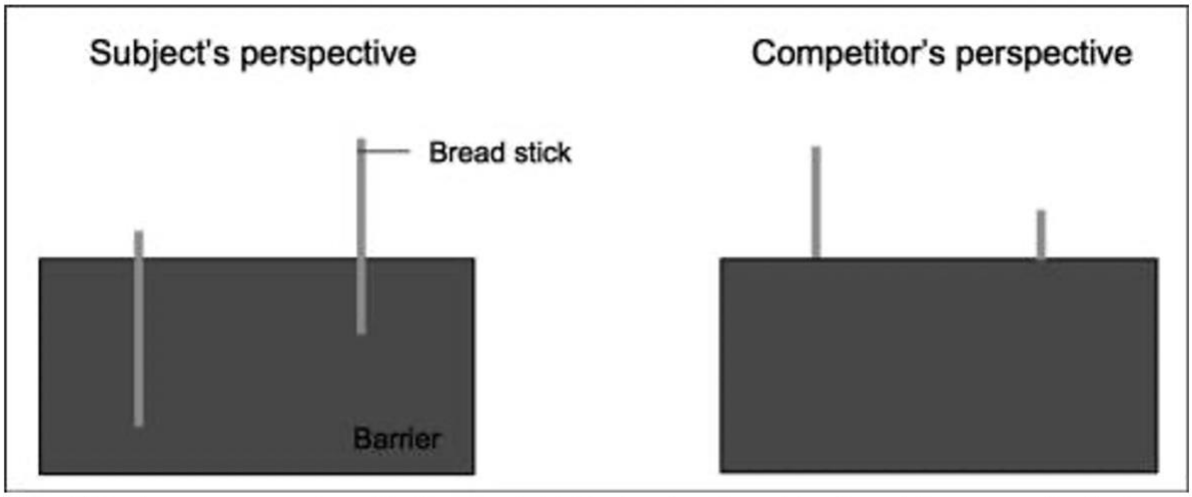
4.1 Level-2 perspective taking

Let’s first consider level-2 perspective taking. As I have explained earlier, level-2 perspective taking concerns about the ability to track how others see things. Level-2 perspective taking ability shows up in children at 3 years of age (Moll & Meltzoff, 2011) or even later, around age of 4-5 (Flavell et al., 1980; Pillow & Flavell, 1986). Evidence from humans suggests that level-2 perspective taking is not easy for children. This is not very surprising since the level-2 perspective taking shares many aspects with a false belief test. To perform a level-2 perspective taking, the agent needs to represent how they themselves perceive the object and also how other agent perceives the same object but from a different angle. Moreover, in certain situations, their own perspective can be different from others’ perspectives. All of these things are also involved in a

false belief task. It is not surprising that there are only a few studies targeting this topic in animals. One of them taking was done by Karg et al. (2016).

In their experiment, the subject chimpanzees compete with conspecifics over two bread sticks when they sat face to face. Between them, a barrier was placed on the table. Two bread sticks of equal size were fixed to the barrier. During the experiments, the bread sticks were placed as in Fig 2 below.

Figure 2



Note. This picture was produced by Karg et al. 2016, showing the subject's and the competitor's views of the bread stick from their own perspective. From "Differing views: Can chimpanzees do Level 2 perspective-taking," by K. Karg, M. Schmelz, J. Call, and M. Tomasello, 2016, *Animal Cognition*, 19(3), p557 (https://doi.org/10.1007/s10071-016-0956-7)

From the subject's perspective, these two sticks are the same size. But from the competitor's perspective, one of them looks longer. In the social condition, subject chimpanzees have learned through experience that they can only get the remaining stick after their competitors' choice. If they chose the same stick chosen by the competitor, they do not get the bread stick. But in the test phase, the subject cannot see the choices of their competitors. After the competitor's turn, both sticks were still there. So, they have to guess which stick the competitor has chosen. If the subject can think from the competitor's perspective, they should avoid the one that looks bigger to the competitor and choose the one that looks smaller from that perspective. In the non-social condition, no competitors were presented. So, the subject shouldn't have a preference between the sticks.

The results showed that chimpanzees did avoid the seemingly bigger one in the social condition more often than in non-social condition. But in the non-social condition, the chimpanzees preferred the seemingly bigger stick. Such results give a mixed picture about level-2 perspective taking in chimpanzees. It seems like the chimpanzees projected their own preference to others in the test. If subject chimpanzees themselves prefer the seemingly bigger stick, they think the others should also choose the bigger one. So, they avoid choosing it in the social condition. This means that the test is not a true level-2 perspective taking test, which should involve projecting a different perspective to other instead of projecting one's own perspective. The mindreading theory would have no difficulty in explaining such an ability. Attributing their own perception to others is a less demanding version of attributing a different perception.

The minimal theory may have some trouble in this situation. It is not very obvious how encountering works here. If the subject chimpanzees encounter the situation as it is, chimpanzees should have no preference between these two sticks. Based on their own encounterings, the competitors should randomly choice between the sticks. However, if the subject chimpanzees somehow encounter the situation as one stick is bigger than another, then they would avoid the seemingly bigger one. In sum, the situation of level-2 perspective taking needs more data to give a clearer picture about how chimpanzees handle such cases.

4.2 The problem of identity

According to Butterfill and Apperly, a signature limitation of their minimal theory compared to the standard theory of mind (they call a *full-blown theory of mind*) is that the latter enables the agent to represent the identity of objects in false beliefs while their theory cannot. Butterfill and Apperly argue that believe, as a propositional attitude, can represent the same object as different identities. However, registration can only capture the relationship between the object and their location but not the identity of the object. Butterfill and Apperly (2013) cite the example from Frege's famous paper "Sense and Reference" (1948) to illustrate their ideas. For Frege, "the morning star" and "the evening star" both refer the same planet but not share the same meaning. Apperly and Butterfill try to make a similar case between the identity of an object and the object itself. What they mean by identity is almost the same as appearance in the real case. For example,

if a human figure doll has two different faces on each side (call each face as Sally and Sally*), the minimal theory will call Sally and Sally* as two identities of the same doll. Based on such treatment of identity, if an agent wants to track the doll's mental activity, the minimal theory and the mindreading theory will predict differently on how the agent will behave based on her attribution of the doll. From the point of the minimal theory, the agent will treat Sally and Sally* as the same because the minimal theorists don't allow the agent to track identity. However, for Apperly and Butterfill, they think the mindreading account should suggest the agent represent the same doll as different individuals, naming Sally and Sally*. But I do not think Apperly and Butterfill's treatment of identity is convincing. In the example of "the morning star" and "the evening star", people treated them as different stars only because they don't know these two names actually refer the same planet. If they know such fact, they will treat these two names as the same in further inference. The key is whether the subject know the fact that Sally and Sally* are the same doll just like the case of Venus. When the agent knew Sally and Sally* is the same doll but with different faces, the agent will treat Sally and Sally* as the same. Or at least, it is not clear why the agent should treat Sally and Sally* as different person. Moreover, it is also unclear for me why, according to the minimal theory, the agent must register Sally and Sally* as the same object. Since Sally and Sally* have different faces, there is no rule or principle in the minimal theory to ban the agent to register them as different persons. From the appearance, Sally and Sally* are indeed different objects.

In Butterfill and Apperly's paper, they discuss the experiment by Scott and Baillargeon (2010) which raises these issues. I think their explanation of the experiment based on the minimal theory is not very convincing. I will go through this study later in this section. Even if I don't think Butterfill and Apperly give a persuasive argument about the limitation of the minimal theory on identity, but if we take such identity limitation as they suggested, experiments by Scott, Richman and Baillargeon (2015) are still not worked in their favour. In these new experiments, Scott et al. show that infants can track the identity of the objects in a false belief test, which is against Apperly and Butterfill's idea about the identity limitation. Also, I also presented this case in more detail later.

Firstly, consider the experiment conducted by Scott and Baillargeon (2010). In this experiment, there are two visually indistinguishable toy penguins. In addition, one is separable which is suitable for hiding a key inside it, and the other one is inseparable. The aim of this experiment is to test whether the infants can understand that the human demonstrator want to put the key in the separable penguins. Through familiarisation trials, infants notice that the demonstrator is always initially presented with the separable penguin in its disassembled state alongside the intact inseparable penguin. Each time, the demonstrator then puts a key into the disassembled penguin and assembles these two pieces together. In the test trials, infants see the demonstrator presented with the separable penguin displayed in a transparent box in its assembled state and the inseparable one covered by an opaque box (the infants are shown that penguin in the transparent box comes apart while the demonstrator is not present). The results show that the infant is surprised at the choice of the demonstrator reaching for the transparent box. The original interpretation made by Scott and Baillargeon suggests the infants know that the demonstrator has mistaken the visible penguin for the inseparable one and incorrectly infers that the separable penguin is in the invisible box. Thus, the demonstrator does not think the separable penguin and inseparable penguin are identical even if they look the same in some conditions. Butterfill and Apperly think this case is not about identity but about types. They argue instead that the demonstrator can register the type of object. Hence the demonstrator expects to see two types of penguins, assembled and disassembled. Through registering an intact penguin in the transparent box, the demonstrator will infer a separable and disassembled penguin is in the opaque box. The problem here is that it is not clear how the infant could register different types of objects using the principles of the minimal theory. According to minimal theory, the procedure that infant might use in attributing a registration to the demonstrator in the experiment could be as follows:

(1) First, the goal of the demonstrator is to put the key into a disassembled penguin. The demonstrator registers <separable penguin, a visible place(either in the left or right side)>

(2) Through the familiarisation trials, the demonstrator will find that the separable penguin is distinguishable from the inseparable penguin from encountering.

(3) In the test trial, the demonstrator encounters an assembled penguin in the visible box. Since the inseparable penguin is the separable penguin according to (2), The demonstrator registers <separable penguin, a visible place(either in the left or right side)>

Thus, the demonstrator will reach for the visible penguin because he registers it as a separable penguin. Following this logic, the minimal theorists hope to explain what happened in the experiment. But such arguments from the minimal theory are flawed. First, principles outlining the theory that I have presented earlier never mentioned anything about attributing representations of types of objects. In this experiment, the infants and the demonstrator know the fact that when assembled, the separable and inseparable penguins look exactly the same. How they register the fact objects that look the same as being of different types is unclear. To introduce the concept of types means introducing more principles at the least. Moreover, employing concepts of types would function much like level-2 perspective taking. The different types of penguins in this experiment can only be represented as different in terms of how they assemble, not by their appearance. Just like the level-2 perspective thanking discussed in the previous section, representing the type of an object involves representing a specific way to perceive the object, here in terms of how the object is assembled. If the minimal theorists take their theory to be able represent types of an object, it effectively allows for level-2 perspective taking. This will contradict the limitations that they take to be characteristic of their own theory. So, the minimal theory has to choose between these two. If the theory keeps the limitation on level-2 perspective taking, it cannot allow for the representation of different types of objects, which means the theory cannot explain the result from the experiment in Scott et al.(2010). If the minimal theory commits to a type tracking ability (let's grant that doing so by patching up the original theory to add such an ability by introducing additional principles), the theory also commits to level-2 perspective taking, which means they need more concepts than encountering and registration. I cannot see how they make this radical a change while keeping their account minimalist.

Scott et al. (2015) conducted a further experiment to demonstrate that 17-month-old infants can track identity. The basic idea of their experiment is to see whether an infant will expect a thief to choose to use a silent toy to replace a rattling toy to steal the rattling toy from its owner. Both the owner and the thief have a preference for the rattling toy. Infants know this through what happens in the familiarisation trials. Since both toys are the same in appearance, the thief wants to dupe owner by exploiting the visual identity of two toys. The crucial point is whether the infant could understand that the thief can cheat the owner through inducing a false belief about identity into owner's mind. In the familiarisation trials, the owner puts either a rattling or a silent toy in front of

the thief, leaving it there and going away. The thief shows interest in the rattling toy, but not the silent toy (she will play with the rattling toy when the owner isn't here). Then the owner comes back, storing if the toy was a rattling puts it in a box. By contrast, silent toys get thrown into a trash bin. Through these trials, the infants should know that both the thief and the owner have a preference for rattling toys. Next, the infants receives either a matching or non-matching test. In the matching, the thief picks a matching silent toy (which is visually identical to the rattling toy in this trial) from the trash bin and replaces the rattling toy left by the owner with the silent toy. While in the non-matching test, the thief replaces the rattling toy with a different silent toy, one that is not identical to the rattling toy but rather is a different colour (it is still similar to the rattling toy in appearance but differs in colour). The results show that the infants notably spent more time looking in the non-matching test than the matching test. This means infants do not expect the thief in the non-matching condition to replace the rattling toy with a different toy from appearance.

According to Scott, Richman and Baillargeon (2015), the infants require a lot of cognitive ability to comprehend the whole process. Infants need to know that both of the participants have the preference of rattling toy and the thief intended to replace the rattling toy with a silent one secretly. The pivotal point is that the thief knows she can deceive the owner on her false belief of the rattling toy. So, the thief must represent that the owner will have a false belief about which toy is there since the owner believes the silent toy is the rattling one. The process involved in the thief's thought process is roughly as follows:

(1) The owner believes that the rattling toy is on the table.

(2) The silent toy is identical to rattling toy in appearance.

If I replace the rattling toy with the silent toy,

(3) The owner will falsely believe that the rattling toy is still on the table.

But according to the minimal theory’s definition of identity, if the infants should represent infants believe the rattling toy is

Switching to minimal theory, it will be very unclear how the theory could explain what happens in the experiment without appealing to these sorts of complex beliefs. I will focus on what the minimal theory could try to say about the false belief about identity. According to the minimal theory, if the rattling toy is apparently identical to the silence toy, they should be registered as the same object. So, the procession of thief’s registration could be as follows:

(1) The owner registers <rattling toy, on the table>.

(2) The silent toy is identical in appearance to rattling toy, so, <rattling toy> is exchangeable with <silent toy>.

If I replace the rattling toy with the silent toy

(3) Therefore, the owner will register <rattling toy, on the table> as well as <silent toy, on the table>.

It is not very clear for me how the minimal theory will predict the infants' expected looking. If infants followed the logic of the minimal theory's understanding of identity, what the thief did is confusing. Because there is no point to change the rattling toy with the silent toy if they mean exactly the same for the owner. Only way out for the minimal theory is to make sure the owner can register the toy can have different appearances. If the infant can represent the owner registers the toy as two toys based on their appearance, it will make sense for the infants why the owner wants to switch these two toys. But just like the previous case of register the type of an object, in original principles, the minimal theory never talks about how register object's appearance. So, the minimal theory will face the exact same problem with tracking object's type. And making things worse, because of this new experiment, the minimal theory has more problem to deal with beside the existing type one.

Summary

The minimal theory of mind account tries to analyse the mindreading abilities of animals and infants in minimal terms. Equipped with their concise set of principles and the concepts of goals, encounterings, and registrations, on this account animals and infants can track others' elementary perceptions and beliefs, without actually attributing perceptions and beliefs. However, a large body of experiments shows that subjects with restricted cognitive resources, such as infants, ravens and chimpanzees, can still track others' beliefs—as beliefs—in a very flexible and broad sense. Any theory of mind which tries to understand mindreading behaviour from such minimal foundations will sooner or later face some phenomenon it cannot explain. Therefore, maybe a better strategy is to work with a full-blown theory of mind at the first place. Chimpanzees, corvids, and infants are likely to possesses a more restricted theory of mind than adult humans, but what they possess is richer than a minimal theory of mind. So, theorists are best off beginning with a full theory of

mind ability and reducing its complexity in response to experimental data as necessary, while maintaining the core elements of a full-blown theory of mind.

Chapter 3 Knowledge attribution theory

In recent years, a number of scholars (Martin & Santos, 2016; Nagel, 2017; Phillips & Norby, 2019; Phillips et al., 2021; Westra & Nagel, 2021) have argued that representing knowledge is more basic, and hence easier than, representing beliefs. They have also argued that since animals have consistently failed to pass false-belief tasks, animals such as nonhuman primates and corvids should be taken to only be capable of tracking the knowledge states of others. I will refer to the resulting theory as the knowledge attribution theory (KAT). By pushing a step further toward attributing mental states like knowledge than the theories discussed in Chapter 1 and Chapter 2, KAT is theoretically more similar to a mindreading theory than those theories. Unlike the previous two theories, it admits that animals can actually represent some of the same sort of mental states that humans represent in mindreading. The key distinction in comparing KAT to the mindreading theory (or the belief attribution theory aka BAT) is that the knowledge attribution theory holds that animals can only represent factive mental states like knowledge and awareness and cannot represent non-factive mental states like belief and thinking. Another argument in support for KAT is that a full-blown mindreading ability is widely considered to be cognitively demanding, arguably beyond the cognitive abilities of animals. And, as many see the evidence, in animal studies, there is abundant evidence showing that animals can track others' knowledge states but fail to pass false belief tasks. Based on the difference between attribution knowledge and attribution belief, the knowledge attribution theory gives an explanation of why that is the case. It is because animals can only represent others' knowledge not their beliefs—and so, in particular, they cannot represent their false beliefs.

In this chapter, I want to argue that knowledge attribution theory is just a true belief attribution theory. And if you are already persuaded by KAT, there are good reasons to adopt the full mindreading. To achieve that, I will first introduce the knowledge attribution theory in detail in the first section. In the second section, I will examine the experimental evidence supporting KAT provided by its advocates and point out that it does not actually work in their favour. In the third section, I will outline the specific limitations to the mindreading theory that are highlighted by KAT, and I will argue that such limitations are not supported by experimental evidence. Finally,

based on all the available evidence, I will argue that there is no reason to prefer KAT to the standard mindreading account.

1. The theory

The main supporters of KAT are Martin and Santos (2016), Nagel (2017, 2021), Phillips (2021). All of them agree that agents are attributing some knowledge states instead of belief states to others during a typical mindreading test. Before I talk about the different between KAT and BAT, I want to make clear the difference between knowledge and knowledge attribution. In Epistemology, knowledge is widely understood to be a factive state that goes beyond true beliefs. Here, the supporters of KAT aren't really concerned with knowledge as epistemologists understand it. What they really want to talk about is knowledge attribution not knowledge itself. But their concept of what is knowledge heavily relied on some widely accepted arguments in epistemology. The most important one is that knowledge is equal to true belief. This is why they claim attribution true belief is not attribution knowledge. However, in practise, KAT simply treats knowledge as a representation of a factive state of the world. Hence, agents can only represent others as knowing things that those agents themselves take to be true. Their inconsistent attitude to knowledge and knowledge attribution gets them into trouble which I will present later in this chapter.

Their core argument lies in the distinction between the knowledge attribution (also known as *factive ToM*) and the belief attribution (also known as *non-factive ToM*). As they suggest, in the field of mindreading studies, for historical and methodological reasons, the primary focus is on belief attribution. They want to shift the focus to knowledge attribution. They argue that knowledge attribution is more basic and easier to perform than belief attribution. By saying more basic, they mean that knowledge attributions are more limited than belief attributions. Knowledge attribution means that the agent can represent others' knowledge states about their surroundings. This ability permits the agent to represent others as knowing or failing to know about their situation in the real world. Unlike belief attribution, knowledge attribution only allows them to represent what they take to be a factual perspective about their environment. In contrast, belief attribution requires the agent to be able to represent both true beliefs and false beliefs. They argue that false belief attribution is the root of the problem in the animal and infant mindreading. This is especially

true in the case of macaques. There is still a lack of evidence to show macaques can pass the false-belief tasks. And it is only recently (Krupenye et al., 2016; Krupenye et al., 2017; Kano et al., 2019) that there have been positive results to support the claim that chimpanzees can pass false-belief tasks. Overall, the experimental data still favours the idea that false-belief attribution is difficult for non-human primates. So, the main advantage of the knowledge attribution theory is to support a good deal of positive evidence of knowledge and perspective attribution in animals without committing the ability to attribute false beliefs.

Moreover, Phillips and Norby (2019) expressed a detailed account about why animals keep failing to pass the false belief task but not the true belief tasks. They argue that there are two core abilities about mindreading: tracking and separation. By tracking, a subject can keep updating how others understand the world. By separation, the subject can keep what other's understand separate from their own understanding. In other words, the ability of tracking enables agents to represent others' understanding of the world (they called it "world map"). While, separating means that agents can have two separated world maps: one is about how others understand the world, the other one is about how themselves understand the world. Only agents that have both capacities are qualified to have the ability of mindreading.

Following this way of presenting the basis of theory of mind, they argue that knowledge attribution meets both requirements of mindreading (tracking and separation). They also claimed that the false belief attribution is the most difficult type of mindreading. In a false-belief task, a participant not only requires the two core abilities of mindreading (tracking and separation) but also needs something more, which is the ability to hold in their mind two contradictory representations. It requires agents to hold both their own (true) representation of the world and their tracking of others' false representation of the world. These representations contradict one another in the false-belief context. To say an agent cannot hold contradictory representations does not mean agents cannot hold two separated representations. Indeed, knowledge attribution theory claims that agents can not only represent other's knowledge but, crucially, also their ignorance. Here, representing ignorance is a special case. Representing ignorance is different from the absence of a belief which not a representation at all. The individual is representing something different from false beliefs and the absence of a belief. This is because in the case of ignorance, the attributers can represent

other's understanding of the world as different from their own understanding. It includes two situations of ignorance attribution. You can either attribute to others knowledge about the world more than you or less than you. In both situations, the attributions do not hold contradict representations. For instance, when someone places a banana in your sight but out of the view of others, you can represent others as not knowing where the banana is, but represent yourself as knowing the place of the banana. Similarly, when someone places a banana out of your sight but in the view of others, you can represent others as knowing where the banana ism but represent yourself as not knowing where it is.

Among the supporters of KAT, Martin and Santos used a different term to refer the knowledge attribution theory. They call it *awareness attribution*. According to them, awareness attribution means that animals can represent other's awareness relations with the world. This is actually even more basic than knowledge attribution because it only concerns one type of mental states: awareness relations. They define awareness relations as connections between an agent and a particular piece of information that the agent knows about world. For example, when your friend sees an apple on the table, you may notice there is an awareness relation between your friend and a portion of the world (namely, the apple on the table). Based on such relations, you can predict how your friend will behave. I treat this awareness relation account as a specific version of KAT. My reason for this is that awareness attribution and knowledge attribution are not fundamentally different. Such idea is also expressed by Santos (2019). In many situations, the content of knowledge attribution is the awareness of others. I will talk more about the difference in later chapters. The most important difference between the theories of these theorists lies on how animals represent others' ignorance. The KAT claims that if the attributer did not know what others know, she or he would ascribe ignorance to others. However, the awareness attribution account argues that agents will represent others' ignorance with more details. One such detail could be based on how they considered others' previous awareness.

2. How it works in animal cases

The knowledge attribution theory claims that it applies in both human beings and some animals. In human cases, it claims that knowledge attribution is acquired earlier than belief attribution, and

that preverbal infants mainly use knowledge attribution instead of belief attribution. My main focus is on animal cases. KAT is largely focused on non-human primates in relation to animals, and in particular apes and monkeys (particularly macaques). Most of the studies on animals mentioned by the supporters of the theory are about monkeys. This is mostly because there is hardly any evidence showing that monkeys can pass the false-belief task. So far, there is only one published study (Hayashi et al., 2020) arguing that monkeys can pass such tasks. Besides this one paper, there is a great deal of experimental work arguing that monkeys cannot track others' false beliefs.

In this section, I will talk about the evidence presented by Phillips and colleagues (2021). They summarised three types of evidence to support the theory. The first type is that non-human primates can represent egocentric ignorance, which means the ability to represent others knowing something we ourselves do not know. The second type of evidence is that, in some cases (Kaminski et al., 2008; Horschler et al., 2019; Horschler et al., 2021), chimpanzees and monkeys cannot represent others' true beliefs while they can represent others' knowledge. The third type is that the knowledge attribution ability in animals exhibits cross-modality, which means non-human primates can make knowledge attributions based on not only visual but also auditory information.

I will classify these three types of evidence into two categories, I call the first category, *general evidence,* and it includes both their first and third types of evidence. I will argue that this category of evidence can be explained by both knowledge attribution and belief. This evidence doesn't favour knowledge attribution theory over belief attribution theory (BAT). The other category of evidence I call *special evidence*. They hold that this category of evidence can only be explained by the knowledge attribution theory. However, I will argue that this special evidence cannot be explained by KAT. If you want to choose a theory to explain it, BAT is still the better choice between them.

2.1 The general evidence

The first category, general evidence, includes the first and third types of evidence provided by Phillips et al(2021). It also includes evidence about the traditional mindreading experiments in animals which are mainly considered as the cases of belief attribution. Since knowledge is taken

to involve factive states, it is no surprise that representing others' knowledge can be understood as involving representing others' true beliefs. In the following paragraphs, I will explain in more detail why general evidence can be explained in the terms of BAT.

The first type of evidence is about representing egocentric ignorance. Phillips et al. used a study by Krachun et al. (2009) to support their point. In the original experiment, two types of apes, chimps and bonobos, were placed in a situation in which they competed with a human for the food that was located between them and the human. The food was placed in one of the two containers when the human competitors can see where the food was and apes can only see that their competitors knew the where the food is. Then, the positions of the two containers were switched in sight of both human competitors and apes. The apes showed a clear preference for the container their competitors reached for. Importantly, a control condition showed that apes will ignore their competitors' choice when they themselves were shown where the food actually is. Based on this experiment, Phillips et al. claimed that primates can represent egocentric ignorance. This is the only experimental evidence mentioned by them to support their claim. Their emphasising of the egocentric ignorance case is because they think it highlights animals' ability to hold two separate representations of the world in their mind at the same time. They think this is more demanding than the normal knowledge attribution in which animals only need to hold one representation of the world. More importantly, the attributers need to represent others know more than themselves about the ongoing situation. But this experimental study can be easily explained in the term of true belief attribution, and this is how it is explained in the original paper. Krachun et al. (2009) called the tests discussed above as "true belief tests". For the belief attribution theory, the behaviour of apes can be explained as followed: the apes followed the choice of their human competitors because they can represent that the human truly believed the foo was there. At least in this case, there is not much difference between representing others' true beliefs and representing others' knowing something. So, this kind of evidence cannot rule out the belief attribution theory. Philips and his colleagues argue that in a slightly different false-belief condition, non-human primates failed to track other's false beliefs. I do not think that the failure of a false-belief task should push people to admit that the previous result must be explained in the term of knowledge attribution as opposed to belief attribution. It is perfectly fine to accept that animals can represent other's true beliefs in the previous situation and fail to represent other's false beliefs in the latter situation.

For the other type of general evidence that Phillips and the colleagues discuss, they argue that since knowledge is not modality-specific (individuals can know things via different sensor systems, like vision and audition), the attribution of knowledge should also be cross-modal. The previous two theories, the behaviour reading theory and the minimal theory, both suffered from problems concerning how to adjust their theory to work with cross-modal attributions such as those found in the current studies. For example, in Santos et al. (2006), rhesus macaques were presented with the opportunity to steal grapes from two containers, one of which will make noise when the monkeys moved it to take the grapes and the other of which will not. When a human competitor was looking away from containers, the monkeys had the choice to steal food from both containers. The results showed that monkeys prefer the silent container over the noisy one. A similar result was found in chimpanzees as well(Melis et al., 2006). These experiments showed that non-human primates can represent others' auditory knowledge to help them decide which actions to take. Again, the evidence that primates can represent others' knowledge through different sensory modal can also be explained by the belief attribution theory. This is because that belief attribution is also not modality specific. What primates achieved in these experiments can be understood in terms of their representing what their competitors believe based on the auditory information. This kind of phenomenon of cross-modal representation is also found in corvids as I discussed in previous chapters.

There are also many further experimental results that are taken as supporting the idea of knowledge attribution. But they also fail to distinguish between true belief attribution and knowledge attribution. Indeed, these studies can be taken as equally supporting both theories. As we can see, trying to separate evidence for KAT and BAT is very hard, especially in the case of tracking true beliefs or knowledge states. In most experiments, researchers designed their protocols to detect different predictions of a mindreading theory versus a non-mindreading theory. Since both KAT and BAT agree that animals can attribute mental states, and so are both versions of mindreading theories, most studies just cannot be considered as evidence solely support one of them. In sum, the first category of evidence, general evidence, for the knowledge attribution theory can equally be used as the evidence for the belief attribution theory. So, it is better to focus on the second type of evidence which the supporters of KAT claim can't be explained by BAT.

2.2 The specific evidence

According to KAT, the specific evidence concerns cases where animals cannot represent others' true beliefs but do not have the problem of representing others' knowledge. In other words, animals can represent knowledge without representing true beliefs. If this is true, it will raise problems for the belief attribution theory. The supporters of the knowledge attribution theory also call such cases as Gettier cases (Phillips et al., 2021; Westra & Nagel, 2021). My main focus is on the category of special evidence because this kind of evidence is what really matters in the debate between these two theories.

Phillips et al. (2021) argue that two studies (Horschler et al., 2019; Kaminski et al., 2008) support knowledge attribution but not true belief attribution. They claim these cases work in similar way as Gettier cases. I think the reason they refer to them as Gettier cases is because, with Gettier cases, these cases are taken to involve true belief that is not knowledge. I will take an in-depth look into these two experiments and show how they work. I will argue that neither of them works in favour of the knowledge attribution, and that these cases are not Gettier cases.

2.2.1 The case of Kaminski et al.

In the study by Kaminski and her colleague (2008), they set up a competitive situation between two agents, one of which was a human competitor, the other one is the experiment subject (where there were three types of experimental subjects: chimpanzees, 3-year-old children, and 6-year-old children). The experimenter placed three opaque buckets between the subject and the competitor, which both agents could take turns to choose from and receive the contents in the chosen bucket. The competitors always chose first and subject individuals' view was blocked while the competitor chose. Next, the subjects had their own chance to choose a bucket. If the subjects want to get the target item (which is a high value item), they need to guess what choice their competitor made.

Experimenters placed a high-quality item in one of the three buckets in the sight of both the subject and the competitor. There were four different conditions. In the *known lift* and the *known shift* conditions, both individuals, the competitor and the subject, will have visual access to the

manipulations. The difference is that in *known lift* condition, the experimenter lifted the food and placed it back in the original location. In *known shift* condition, the experimenter changed the location of the food in a different location. In both conditions, both agents have the same information about where the target item is. The only difference is that the *lift* condition should be less demanding in terms of memory than the *shift* condition. This is because in the *lift* condition, the subject only needs remember the original location of the food. However, in the *shift* condition, the subject needs to update a new location.

In the two remaining experimental conditions, the *unknown lift* and the *unknown shift* conditions, only the experimental subject has visual access to the manipulation while the competitor's view is blocked. The manipulations in these two situations were otherwise the same in *known lift* and *unknown shift* conditions. In search of these different conditions, the experimental subjects have the chance to choose the bucket only after their competitors chooses. In the two *unknown* conditions, the experimental subject has some additional information compared to their competitors. If the subjects can attribute either knowledge or belief to others, they should behave accordingly based on their additional information.

Results showed that chimpanzee subjects chose the bucket in which they last saw the high-quality food placed in more often when their competitors had not seen the final manipulation. This means that chimpanzees behaved the same in the conditions of *unknown lift* and *unknown shift*. They preferred to choose the bucket which actually had the food in it no matter whether the food had been lifted or shifted by experimenters. But these conditions have a big difference. Since the *unknown lift* condition procedure did not really change the location of the food, the competitor should still have a true belief where the food was. However, the procedure of the *unknown shift* condition changed the location of the food. So, competitors should have a false belief of where the food was. In the experiment, only the group of 6 years old children acted like they could represent others' false beliefs.

Based on this study, Nagel claimed that "nonhuman primates have difficulty representing true belief that falls short of knowledge" (Nagel, 2017, p. 534). Phillips et al. argued that "apes and

monkeys fail to represent others' true beliefs in cases where they have no trouble representing others' knowledge" (Phillips et al., 2021, p. 6).

I will carefully unfold the logic in argument here from supporters of the knowledge attribution theory in the following. The knowledge attribution theory can explain the result in this experiment in the term of knowledge and ignorance attribution as follows (focusing on the chimpanzee subjects):

1. Both the *unknown lift* and *unknown shift* conditions are considered as ignorance conditions because in both cases both the competitor and the chimpanzee subjects did not know what happened (that is, following the manipulation they were ignorant regarding the location of the food item).
2. The different performance of the chimpanzee subjects in known and unknown conditions suggests that they can represent their competitors' mental states as knowledge and ignorance.

On the other hand, the belief attribution theory will have difficulty in explaining what happened with the belief attribution:

1. The *unknown lift* condition involves true belief attribution because the competitors still hold a true belief of where the food is. Hence, subjects should represent that the competitor has a true belief. The *unknown shift* condition involves false belief attribution because competitors hold a false belief about the location of the food. Hence, subjects should represent competitors' as having a false belief.
2. Based on true belief attribution, subjects in the *unknown lift* condition should behave like they do in the *known* conditions, which means subjects should not reach for the high-quality food. This is because they should assume that the competitor, who they represent as having a true belief, will have chosen this bucket already. However, contrary to the belief attribution theory's prediction, the chimpanzee subjects still choose the bucket that has the high-quality food items. So, it seems that the chimpanzee subjects do not represent others' true beliefs in this case.
3. The same applied in the *unknown shift* condition. If the subjects can represent others' false beliefs, they should take advantage of that to get the high-quality food comparing to the

unknown lift condition. But subjects did not show that preference in the tests. So, they did not represent others' false beliefs in this condition.

This interpretation of this experimental data is not watertight. In this study, researchers did the same test across three groups. The comparison between the chimpanzees and the 6 years old children is particularly interesting. Children over 6 years are undoubtedly considered as having the full-blown mindreading ability. They did show the false belief attribution ability in the test of *unknown shift*. Children chose the bucket with the high-quality reward more often in the *unknown shift* condition than in the *unknown lift* condition. However, the results also showed that children, like chimpanzees, chose the bucket that they knew had the reward in it in both unknown conditions, in comparison to the known conditions. The difference with chimpanzees is that children over 6 showed an extra comprehension of the false belief condition. If children represented others' true beliefs in the *unknown lift* condition, they should not choose the bucket that the reward was placed in. The fact that both chimpanzees and children chose the bucket the high-quality reward was placed in in the *unknown lift* condition more than in the *known lift* condition cannot be understood in terms of their not representing others' true beliefs. Instead, it seems as though it should be explained by other beliefs they hold. For example, blocking the sight of their competitors will give the subject the belief that their competitors are not likely to be as confident about the location of the reward as they are in the *known lift* condition. So, even if they could understand their competitors' as having true beliefs, they will still try to take advantage of the fact that their lack of confidence in unknown conditions might have made them choose the wrong bucket. When comparing to the performance of 6-year-old children, the result of this study should be more considered as the proof of chimpanzee's inability in false belief tasks comparing to 6-year-old children.

2.2.2 The case of Horschler et al.

The study by Horschler and the colleagues (2019) was designed to follow up on the study of Kaminski et al. They wanted to dig deeper into the *unknown lift* condition to see what exactly affected the subject's choice. In their study, they used the expectancy violation paradigm instead of evidence based on direct reaching used in Kaminski et al.'s study. In the expectancy violation paradigm, experimenters measure the looking times of experimental subjects as their prior

predictions about what happened. The basic assumption is that the longer the subject looks at a scenario the more surprising they are. So, the looking time of the subject indirectly suggests what the subject doesn't expect to happen.

In the experiment, rhesus macaques were placed opposite to a human agent. They both watched a piece of fruit being hidden in one of two boxes. Then the human agent's view was blocked by a an occluder. Then monkeys were put into two conditions. In one condition, the *fruit moves* condition, the fruit moved out of the box and then back into the same box, but only the monkeys can see the move, the humans can't. In the other condition, the *box moves* condition, the box that contained the food is flipped open and then closed, and again, only the monkey can see this happen with the box the human can't. In both conditions, the fruit ends up in the box it started in and the events that happen are only visible to monkeys. The difference is the fruit didn't move at all during the *box moves* condition. This design is based on the assumption that monkeys pay attention to the target item (here this is the fruit), and that updating this information creates an additional cognitive load for the monkeys. So, in *box moves* condition, the monkey should have no trouble to remember where the fruit is, while monkeys in *fruit moves* condition (where they have to update the location twice), they may be confused about the location of the fruit. In both conditions, the occluder was dropped after the manipulations above, and then the human agent reached into one of boxes while the monkeys looked on. Based on monkey's information of where the fruit is, one of the boxes is the correct one which has the fruit inside. Meanwhile, the looking time of monkeys for both correct and incorrect reaching were recorded. The idea is that if the human agent reaches for the correct box as monkeys think they should, monkeys would not be surprised by such move. But if the agent reaches for the incorrect location, monkeys should be surprised.

The result showed that in the *fruit moves* condition, monkeys showed no significant difference in looking time between the correct and incorrect reaching. This means that monkeys did not have any particular expectation about the human agent's behaviour. By contrast, in the *box moves* condition, monkeys looked significantly longer when they saw an incorrect reaching than a correct reaching. This suggests that the monkeys did have an expectation that the human agent should reach the correct location of the fruit in this condition. Together, these results could be explained

by supposing that the monkeys expect their human competitor to still know where the food is in the *box moves* condition but not in the *fruit moves* condition.

Phillips et al. (2021) treat this study as being much the same as the Kaminski et al. study. They think it showed that monkeys can represent others' knowledge but not true beliefs. However, I will argue that this study is even more problematic for the knowledge attribution theory.

In this study, both the *fruit moves* and the *box moves* conditions should be considered true belief attribution conditions. In both conditions, the human agent should still hold a true belief about the location of the fruit. Here's how I see KAT supporters using this experiment to argue against BAT:

Monkeys should represent that their human competitors have a true belief about where the fruit is. So, they should make the same prediction about which box the human will reach for. However, the results showed that the monkeys in the *fruit moves* condition did not predict that the human possessed a true belief of the location of the fruit. This shows the that monkeys failed to represent others' true beliefs in this condition. Meanwhile, in the *box moves* condition, the monkeys had no trouble representing others' true beliefs. According to KAT, in the *fruit moves* condition, monkeys represented the human was in a state of ignorance, while in the *box moves* condition, the monkeys represented the human's knowledge state.

Together, with the results from the Kaminski et al. study, advocates of KAT take this case to show that "apes and monkeys fail to represent others' true beliefs in cases where they have no trouble representing others' knowledge" (Phillips et al., 2021, p. 6).

Putting this study and the previous study by Kaminski et al. together, however, show the inconsistency in the treatment of ignorance by advocates of KAT across the two studies. In Kaminski et al.'s study, any situation in which only the subject knows what happened to the target object, including both unknown lifting and unknown shifting is taken to create a state of ignorance in the competitor. But in the study of Horschler et al., only a specific way of handling of the object outside the view of the competitor is considered as creating a state of ignorance in others. More specifically, moving the box is not considered to create ignorance in others, while moving the fruit

is taken to trigger others' ignorance about the object. Without this different way of understanding what leads to ignorance, advocates of KAT cannot explain the different results in the two conditions in this study, since on the understanding of ignorance used in the Kaminski et al. study, both conditions in the Horschler et al. study would count as ignorance conditions and so should lead to the same results for the two conditions. The only alternative would be to treat the conditions of moving the box and that of moving the fruit as involving two different kinds of ignorance. But it is very hard to imagine how the subject can represent two different types of ignorance of others. According to the theory of knowledge attribution, individuals only represent ignorance when others do not know something. In other words, when an agent represents others as ignorant, the agent does not distinguish what it is that makes others fail to know. As long as agents represent others as ignorant, they should behave the same. So, the knowledge attribution theory should predict that subject monkeys will expect their competitor to behave the same in both the *box moves* and the *fruit moves* conditions.

This problem undermines KAT as I have presented it. But perhaps an alternative version of KAT can get around this objection? As I have mentioned previously, Martin and Santos (2016) provide a specific version of KAT, which they call an *awareness relation account*. On their account, agents can attribute awareness relations between other agents and information about the world. This ability enables animals to know whether others are aware of the things the animals themselves aware of. Moreover, Martin and Santos propose a detailed feature about awareness relations. They claim that awareness relations will be disrupted by any manipulation of the object's location when this happens outside the awareness of the agent. Due to this specific feature of awareness, ignorance attribution will be different from general KAT. For general KAT, any manipulation unknown to an agent will be represented as ignorance. For the awareness relation account, only if the unknown manipulation involves the object's location will it be represented as ignorance. Returning to the experiment conditions, notice that both the *fruit moves* and the *box moves* conditions lead the primates to represent their competitors as ignorant. However, only in the *fruit moves* condition does the location of the food change. In this condition, monkeys should fail to represent others' awareness relations to the food. This is why monkeys failed to track others' knowing in the *fruit moves* condition. In *box moves* condition, monkeys still can represent there being an awareness relation between the competitor and the food. This is why monkeys can track

others' knowing in the *box moves* condition. Another Horschler et al. (Horschler et al., 2021) experiment might be taken to support the awareness relations version of KAT by emphasising the importance of location in awareness relations.

To this awareness relation account, I will say it is more like of an ad hoc hypothesis. The results of the experiment can be explained by other hypotheses other than the awareness relations. For example, the *fruit moves* condition is more demanding for monkeys' cognitive capacities compared to the *box moves* condition. So, monkeys confused about where the fruit is in the *fruit moves* condition. This will explain why monkeys think that the human agent doesn't know where the fruit is. Even if such hypotheses can explain what happened in this study, I wouldn't claim this study is a case supporting BAT instead of KAT. It is better to do more further studies to check how these hypotheses hold in a designed setup.

The evidence here is interesting. But I don't think it can be explained better by KAT compared to BAT. One thing that can be concluded is that the original KAT proposed by Phillips et al. (2021) cannot explain what happened in these studies. Only the awareness relation account (Martin & Santos, 2016), with some additional hypotheses beyond the original KAT suggested, can explain the data. But BAT can do as good as the awareness relation account without such hypotheses. In sum, by examining all the evidence provided by KAT, there isn't a single case only which can only be explained by KAT instead of BAT.

2.2.3 Are these cases really Gettier cases?

At first glance, the cases involved in both experiments seem like Gettier cases. But they are not. In a typical Gettier case, some individuals have a belief which is both true and justified, but which still fails to be knowledge. In both experiments, the human competitors only have a true belief but not a justified one. They only know where the food is at the beginning, but they cannot guarantee the food will still be in the same place when their sight is blocked. For example, the *box moves* condition in Horschler et al. (2019)'s experiment, human competitors will have visual access to the food's location in the beginning, and then their sight will be blocked. Out of their sight, the box (which covered the food) is opened and then closed. If this situation counts as a Gettier case, the animal subject should think the human competitor holds a justified true belief about where the

food is, but it fails to be knowledge. From the perspective of the human competitor, s/he may think that the food is still in the original position after the manipulation, but this view cannot be a justified one. She or he may still prefer to think the food is in the original place since there is no evidence to show it is no longer there, but this is not strong enough to justify that belief. If the human competitor does not have a justified true belief in this situation, it cannot be a Gettier case.

Why do the supporters of KAT want to call these situations Gettier cases in the first place? I suggest that they want to guide readers into a tradition philosophical debate which knowledge is different from true belief. If readers agreed with them that these experiments are Gettier cases, they would automatically think of knowledge as something different than justified true belief, and so could perhaps have knowledge without belief. As I have argued in the previous paragraph, that is not the case. For example, in both *box moves* and *fruit moves* condition, the human competitor did not have a justified true belief about where the food is. Thus, there is no guarantee how they should behave based on an unjustified true belief. Even if the subject monkey can attribute belief to others, they still do not have a clear prediction of how others should behave under these conditions. To explain why in tests monkeys made different prediction in these two conditions is another matter. But as I have argued, there is no reason to suppose that the different prediction is due to the monkey's only having the ability to represent knowledge or ignorance, and not belief.

3. The limitations of knowledge attribution theory

Since KAT advocates that representing the knowledge state is more basic than represent the belief state, it is reasonable to think that the knowledge attribution theory will take animals to have some limitations compared what a belief attribution theory would say. Such limitations manifest in how each theory predicts what animal can do under their theories. In this section, I will focus on the proposal from Westra and Nagel (2021). They suggested KAT would predict animals will show limitations in two situations: level 2 perspective-taking and representing ignorance. In the following sections, I will discuss these two hypothesized limitations in detail and point out that such limitations are compatible with either theory, and more importantly, only the BAT is compatible with the absence of the limitation.

3.1 Level 2 perspective-taking

The first limitation is about level 2 perspective-taking. They express it as follows: “factive mindreading does not capture the fact that a person's knowledge of the world might be represented in a particular way that might differ from our own — what some refer to as “Level 2 perspective-taking” (Westra & Nagel, 2021, p. 6)

Level 2 perspective-taking is about the ability to understand how others perceive the world from their specific viewpoint. It is different from level 1 perspective taking, which refers to the ability to understand the content others see in a situation. Level 2 perspective-taking is considered more cognitively demanding, because it not only requires the agent to understand what others can see but also how they see it. According to the knowledge attribution theory, others’ knowledge of the world is based on the attributer’s knowledge of the same world. So, the attributer can represent others’ knowledge based on the fact that they share the same world. It also means that the way the attributers represent others’ knowledge is the same way they themselves represent the world. They cannot represent others’ perception through their eyes. However, the ability to represent the way others’ are perceiving is exactly what is required by level 2 perspective-taking. As I have noted, Westra and Nagel rightly pointed out that the knowledge attribution should not allow attributers to do level 2 perspective-taking.

How does a belief attribution theory demand the ability of level 2 perspective-taking? In general, a belief attribution theory also does not require the level 2 perspective-taking. The ability to represent others’ true or false beliefs does not tell much about how they represent others’ beliefs. Level 2 perspective-taking is a very demanding process even for human children. The current evidence (e.g., Flavell et al. 1981; Frick et al. 2014) shows that it not until children are around 4 to 5 years old that they start to show the level 2 ability. So, it would not be surprising to find out that most animals did not show level 2 ability in the test. Even considering the typical false belief tasks done with infants, these do not require the agent to do level 2 perspective-taking. In the debate between knowledge attribution and belief attribution, level 1 perspective-taking is much more important. The level 1 ability is required to track others’ understanding of the world, but the level 2 ability is a luxury requirement.

There is only one piece of evidence in animal studies showing that animals might do level-2 perspective taking. As I have discussed in the previous chapter, Karg et al. (2016) demonstrated that chimpanzees can do a task requiring level 2 perspective-taking. If animals can do level-2 perspective taking, this means KAT cannot explain such behaviour. But as I have presented this study in chapter 2, it isn't that clear whether chimpanzees can actually do level 2 perspective-taking from current study. But at least, the available data favours BAT on this matter, since BAT is compatible with animal having a level-2 perspective taking ability but KAT is not compatible with this.

3.2 Ignorance

The second proposed limitation is about the way individuals represent others' ignorance. Westra and Nagel claim: "It also does not distinguish between various ways of failing to know, between ignorance that consists in simply lacking a belief on a certain point, having a false belief, or having a true belief that fails to rise to the level of knowledge." (Westra & Nagel, 2021, p. 6)

As they present it, others' ignorance could be due to multiple causes. Individuals will only represent others' ignorance in one way, even if the causes of their ignorance are different. But I will argue that this limitation suggested by KAT is not supported by current studies. Before I present the counter arguments, let's focus on what ignorance is and where different types of ignorance come from.

Ignorance is the opposite of knowledge, referring to agents failing to know something. According to Westra and Nagel (2021), it includes three types of ignorance: true beliefs that failed to be considered as knowledge, false beliefs, and the absence of any beliefs. Contrary to their views about ignorance, others (Phillips & Norby, 2019; Phillips et al., 2021) talked about two different types of ignorance: egocentric and altercentric ignorance. Phillips and Norby (2019, p. 15) write that "there are still two different ways in which others' maps could differ from yours". Altercentric ignorance means that you represent others as ignoring something you know, and egocentric ignorance means that you represent others as knowing something you ignored. This way of classifying ignorance attribution is confusing compared to the way Westra and Nagel suggested. The concept of egocentric ignorance is not a typical form of ignorance we talked about, because

in this situation, you do not represent others as ignorant at all. What you are representing is that someone knows something I do not. This is about representing others' knowledge even if you do not know what the content of that representation. When we talk about ignorance attribution, it should be about individuals representing other's failing to know a certain situation independent of their own knowledge about such situation. Altercentric ignorance is only one kind of ignorance, because there could also be situations where both attributers and attributees do not know something.

On Westra and Nagel's account, the second limitation of the knowledge attribution theory is that individuals can only represent that others do not know something but not why they do not know. The reason for this limitation, they suggested, is that the ability to attribute knowledge only enables individuals to represent others' mental states matching with the reality. When others' mental states are different from reality no matter how, this non-factive representation will demand much more cognitive capacity than the factive representation. This limitation is large due to the fact that the knowledge attribution easier to achieve compared to the belief attribution. If primates really had this limitation of ignorance representation, how should they behave? Most of the situation we have discussed are about setups involving competition for food. Naturally, when you think your competitor does not know where the food is, you should predict they will choose randomly. So, this means that if this limitation of ignorance attribution is right, animals should predict competitors should behave the same (choosing randomly) in false belief cases and true belief cases that are not knowledge cases. The results from false belief tasks in primates showed a mixed picture. Consistent of the forecast with KAT, Marticorena et al. (2011) showed that monkeys made no prediction about how a human competitor will act when she holds a false belief. But in opposition with that theory's prediction, Kaminski at el. (2008) showed that chimpanzees will predict a human competitor will act based her true belief when she actually holds a false belief.

4. The problem with KAT

As I have mentioned in the previous sections, I plan to argue against KAT from two levels. First, the experimental evidence does not favour a knowledge attribution theory. This claim involves two parts: (1) the evidence provided by the knowledge attribution theory supporters does not work in their favour, and (2) there is evidence that cannot be explained in terms of knowledge attribution.

Second, I have argued that KAT can be reduced to a belief attribution theory with a twist. For the first part, regarding the experimental evidence, as I have argued in the previous sections, the experimental evidence does not favour a knowledge attribution theory over a belief attribution theory. In the following, I will focus on my second argument against KAT. In short, KAT is a redundant theory.

The main reason to support a knowledge attribution theory is based on animals' poor performance in false-belief tasks. By ruling out the non-factive part, which includes the false beliefs, the theory claims that non-human primates can only represent others' knowledge. But I think what happened in infant studies can give us some enlightenments. The standard version of false belief test for children, like the Anne-Sally test , has been shown to require more than theory of mind ability. Things like linguistic ability and executive control are also necessary. By reducing the involvement of these additional cognitive demands, infants were able to pass adjusted versions of false belief tests. For example, by reducing processing demands in the test, Setoh et al. (2016) showed that 30 months old children can pass such test. A similar claim can be made in animals' cases. Animals' failure in many false-belief tasks is due to other factors, such as the setup in experiments or other limitations in animals' cognitive capabilities. Considering recent positive results from experiments on false belief tasks (Krupenye et al., 2016; Buttelmann et al., 2017; Kano et al., 2019; Hayashi et al., 2020) this suggests a more promising picture for the belief attribution theory.

If we address the still controversial debate about false-belief debate in non-human primates, is there still a solid argument to prefer the knowledge attribution instead of the belief attribution?
To answer this question, we should step back first. What is knowledge in the context of this debate about animal's mind, and how does it relate to belief? Here are two ways to think about the relation between knowledge and belief.

4.1 Knowledge with belief

On the first view, knowledge is based on belief. Within epistemology, most philosophers would agree that you can only know what you believe. For example, it will be very bizarre to claim you do not believe some fruit is in the refrigerator, but you know that. In order to know something, you must first believe it. Based on this understanding of knowledge, let's think a step further—what is

involved in tracking other's knowledge? Similarly, there should be two ways: by tracking others' beliefs or not.

If you are tracking beliefs, this means that when you want to track what others know, you just track what they believe. By comparing their beliefs with what you yourself believe about the world, you would track what they know about the world. In a typical case, you only need to track others' true beliefs. If we accept this way of understanding knowledge and tracking others' knowledge, there is no need to separate knowledge attribution from (true) belief attribution.

So, KAT has to take the other view here, which is that we track others' knowledge without tracking their beliefs. The problem is how to do that. Given that knowledge should be based on belief, is there a way to do that? KAT suggests that, in tracking others' knowledge, attributers only need to represent the factive world from their own perception. The attributer will first have a world map about the real world, and then represent others' world map based on this map. If we agree any knowledge must build on true beliefs, then the attributer's own understanding of the world, the world map, should also be built on true beliefs. When agents attribute knowledge to others based on their own knowledge, it should also be built on true beliefs. I do not think there is a special way to track other's knowledge without the help of beliefs. It either comes from attributers' own belief or from other's own belief.

In sum, if you agree that knowledge is based on belief, then you have to agree that the knowledge attribution is also based on the belief attribution. It turns out the KAT is a redundant theory.

4.2 Knowledge without belief

The second option is to think the knowledge is not based on belief, in that case what is knowledge? Philosophers like Timothy Williamson (2000) have proposed a knowledge first account. He argues that knowledge is the factive mental state which does not need to further analyse in the terms of belief, justification, and truth. Knowledge already entails such things. For the purpose of my discussion about knowledge and belief, I do not want to extend the debate in the broad background of epistemology. I only want to see whether the knowledge first argument works in animal cases. If we consider this argument in simple and restricted situations like those many cases in animal

studies, it may be more plausible. In a typical animal study, animals' knowledge focuses on their surroundings, and mainly concerns about rewards (in most cases food) and the competitor or collaborator. Their knowledge in these situations focuses on what they can perceive, and where they turned their attention. For example, imagine a situation in which two chimpanzees compete for a piece of food between them. The food is placed in one of two containers which are visible for both chimpanzees. How does the subject chimpanzee track the other's knowledge about the food? The subject chimpanzee only needs to represent what the competitor perceived. The subject does not need to bother about what the other believes. People may argue this situation is too simple to distinguish between knowledge attribution and belief attribution. It could be explained in terms of perspective taking or even behaviour rules. My point here is not to show that only mindreading theory can work but ask whether we can give a case where knowledge attribution can work without belief involved. In a simple situation like the one we have just been considering, knowledge attribution definitely can work without beliefs, but the contrary is also true. This situation could also be explained by belief attribution, without knowledge attribution. The evidence which is supposed to exclusively support KAT over the belief attribution theory is the so-called "Gettier cases". As I have argued in the previous sections: first they are not the Gettier case, second the experimental results actually do not fit the explanation provided by KAT. Apart from this evidence, both the knowledge attribution theory and the belief attribution theory work fine to explain animals' behaviour. Therefore, why not KAT?

We need to talk about the full picture of mindreading not only in animals like non-human primates but also human children. Since knowledge attribution theory only works in so to say simple situations which exclude tasks like false beliefs, how can it explain more complex situations? Advocates of KAT said they use belief attribution to explain these situations. Knowledge attribution theory is already using simplified understanding of knowledge, and it only worked in undemanding situations. We also know that there are situations where knowledge attribution can be more complex and demanding, like understanding a suspect's knowledge in a crime scene. Should we distinguish a simple version of knowledge and a full-blown version of knowledge just like we did in the mindreading debate based on belief attribution? How is a simple version of knowledge different from the simple version of belief in a meaningful way, or are theorists just trying to dodge away from the problems they take to be associated with belief attribution by simply

changing the name of the process involved? For me, the answer is clear. KAT does not provide a substantial difference from belief attribution theory in the animal cases, and even worse in the human children's case because KAT requires the help of belief attribution to explain them.

Both of the two different ways of understanding knowledge (as involving belief, and as independent of belief) lead to the same conclusion—that the knowledge attribution theory is a superfluous theory. What KAT can explain, the belief attribution theory can also explain. What KAT cannot explain like false belief attribution, BAT can explain as well.

Summary

The knowledge attribution theory is by far the closest theory to the standard mindreading theory among the alternative theorise I discuss in this thesis. It agrees with many elements of the mindreading theory or, as I have referred to it in this chapter, the belief attribution theory. Both agree that the attributer can attribute standard mental states to others, something that both the behaviour reading account and the minimal theory of mind account reject. The disagreement between KAT and BAT is subtle but still significant. KAT claims animals can only attribute factive states like knowledge to others. The justification for this claim is mainly based on two arguments: 1) KAT theorists take a handful of evidence from animal studies to support KAT and disapprove BAT; 2) they think animals' miserable record of performance on false belief tests is hard to justify based BAT. In this chapter, I have argued that both of these arguments fail. First, the evidence they provided should not be treated as supporting KAT and, second, there are more and more positive results from recent experiments showing animals *can* pass false belief tests. Without these two arguments to support KAT, in the broader picture, there is really no reason to prefer KAT over BAT. Because most of the studies with animals have been designed to examine the support for BAT. KAT needs their own experimental protocol to find support for its claims. Moreover, the core idea that motivates KAT is the promise that representing factive states is simpler than representing non-factive ones like belief. But this promise is never illustrated in any detail by KAT advocates, much less directly defended by them. So, I can see little reason to switch from the well-defined mindreading theory to a very similar but less defensible knowledge attribution theory.

Chapter 4 ToM in chimpanzees

Chimpanzees are the most studied species in the field of animal mindreading. It makes them the perfect topic in the discussion of ToM in animals. In this chapter, I want to focus on one specific question, which I think is the most important one: do chimpanzees have a theory of mind? To answer this key question, we need to have a basic idea of what a theory of mind is. Then based on the understanding of ToM, we need to figure out whether chimpanzees are capable of having one. I will separate the discussion into two parts. In the first part, I will talk about what defines the ability of ToM and what kinds of findings should count as evidence for ToM. In the second part, I will examine what kinds of mental states chimpanzees can attribute to others by critically examining the experimental evidence. While most philosophical discussions of whether chimpanzees have a ToM focus very narrowly on the question of whether they can understand belief, my discussion will be much broader. So, in this part, I will not only investigate how chimpanzees behave on belief-related tasks but also explore how they perform on tasks required for other mental states like emotions and desires. By enlarging the scope of mental attribution considered, we can see how chimpanzees' mind work in a broader picture. To answer the question whether chimpanzees possess a ToM, I will argue that chimpanzees do have a theory of mind from three different levels: attributing goals and intentions, attributing perception and knowledge states, and attributing beliefs including false beliefs. Based on the current evidence, the most reasonable explanation is chimpanzees possess a ToM. Other alternative theories which I have discussed in the previous three chapters failed to fully explain the behaviour of chimpanzees. One more thing worth mentioning here is how to examine the experimental evidence in the field. I will check each piece of important evidence from the view of other alternative theories.

1. ToM and the evidence

To answer the question of whether chimpanzees have a ToM, we need first to figure out what kind of ToM we are talking about. First, I want to show you how others are using the term ToM and its synonyms. In the landmark paper from Premark and Woodruff, they described an individual possessing a ToM as someone who "imputes mental state[s] to himself and others" and "a system of inferences of this kind is properly viewed as a theory" (Premack & Woodruff, 1978, p. 515).

From then on, a theory of mind normally has been taken to require two separate parts: ascribing mental states to others and a "theory" or system which can reason based on others' mental states. But when it comes to what exactly that means, people have emphasised different parts. For example, Penn and Povinelli talked about their minimal standard of ToM "the system must be able to produce and employ a particular class of information, namely information about the state of these cognitive variables from the perspective of that agent as distinct from the perspective of the system itself" (Penn & Povinelli, 2007b, p. 733). They highlighted the importance of behaviour cues. Heyes emphasised the reasoning ability by saying someone with ToM thinks "mental states play a causal role in generating behaviour and infers the presence of mental states in others by observing their appearance and behaviour" (Heyes, 1998, p. 102). Call and Tomasello took ToM to involve "understand[ing] the psychological functioning of others" (Call & Tomasello, 2008, p. 187). In general, people mostly agree that in having a theory of mind, an individual will be able to understand others' mental states, including their goals, perceptions, beliefs and so on. Many researchers would also agree that a theory of mind should enable an individual to predict the behaviour of others based on their mental states. In sum, ToM refers to a cognitive ability which enables the agent to understand, and potentially predict, other agents by representing their mental activities. Another related question is about how such a ToM capacity is realised. As I have discussed in the previous chapters, the main theories are theory-theory (TT), simulation theory (ST), or a hybrid theory based these two. By analysing experimental data in chimpanzees, we may get some ideas about how chimpanzees achieve ToM comparing among TT, ST, and hybrid theories.

To understand what evidence should be counted as supporting ToM, we need to go back to the history a little bit. In current research, passing the false-belief task is widely considered to be a golden standard. But I don't think it is the only important requirement for an individual to have a ToM. Why is the false-belief test so important in the literature of ToM? It is largely driven by the worry that in many situations it is very hard to distinguish simple behaviour reading from mindreading. But the false-belief task is not immune from attack from such worry. That is exactly what happened in the study of young infants. There is much positive evidence showing that infants below 3 years old can pass nonverbal false belief tasks, and nevertheless the debate is still alive. In the study of chimpanzees, there is also now some evidence to show chimps can pass the

nonverbal false belief task as well(Krupenye et al., 2016; Buttelmann et al., 2017; Kano et al., 2019; Hayashi et al., 2020). Still, some researchers are not satisfied with the results. For example, Horschler and colleagues (Horschler et al., 2020) questioned the method used in those experiments. They argued that methods such as anticipatory looking are hard to repeat in human studies. Putting these disputations aside, how can we decide whether an individual has a ToM? Some scholars (Heyes, 1998; Penn & Povinelli, 2007b; Lurz & Krachun, 2011) want to solve the problem by upgrading the standard. They think that by designing a more rigorous version of the false belief task they can overcome the drawback of previous studies. But I doubt that is the right solution for our current dispute. There is no one-off solution. Just like other scientific theories in history, there is no decisive evidence that can support or overthrow a theory. A better-designed study is helpful in its own way, but it can never be decisive on its own. In sum, I will argue that when we look beyond single experiments at the overall results, that they clearly favour the ToM account over the alternatives.

In general, there are three types of alternative explanations about why chimpanzees can perform mindreading related tasks. Type one is behaviour reading which claims that the ability to attribute certain states to others is based on pure behavioural cues. This is covered in chapter one. The main idea is that instead of ascribing an inner state to another agent, chimpanzees only need to learn the behaviour rules which guided the behaviour of another agent. Type two is a group of theories that lie between pure behaviour reading and mindreading. Most of them do not agree with behaviour reading, but they also do not think chimpanzees have a ToM. One example is the minimal theory of mind (Butterfill & Apperly, 2013) which argued that what chimpanzees attribute to others is not a mental state, rather it can be explained in terms of non-mental mechanism. I have presented their idea in chapter two. Another example of a type two theory is submentalizing, as proposed by Heyes (2014, 2017) which claim that individuals use some domain-general cognitive mechanisms instead of domain-specific mindreading. This idea is contrary to the general idea that ToM is meant to be a domain specific ability, and not the product of domain general learning.

To address the worries relating to the behaviour reading theory (BRT), it is always about experimental setups. As I have argued in chapter one, both theories need to provide their own evidence. Such responsibility is not only on the mindreading theory. Proponents of BRT like Penn

and Povinelli argued that chimpanzees used behaviour cues to understand others instead of mental states. But in experimental studies, the most obvious kind of behaviour cues can be controlled for and ruled out as the explanation for chimpanzees understanding of other's behaviour. It is impossible to rule out all possible behaviour rules in just one experiment setup. So, as long as the experiment considers certain behaviour rules in its design, it should be taken into consideration in the discussion.

For other alternative theories, one of the concerns is how can we make sure chimpanzees are attributing mental states instead of some non-mental alternatives? For example, the minimal theory of mind claims chimpanzees are using a simpler mechanism like registration instead of beliefs to prediction others' behaviour. Because of their simplification strategy, there should be some testable differences in experimental setups between these two theories.

In general, if we can find the right kind of experimental evidence, worries from other theories should be remitted. The basic requirement for a study to be considered as good evidence for mindreading is that it should contain at least one behaviour rule control setup, which rules out the possibility that chimpanzees can use a simple behaviour cue like pointing to help their decision. The progress is based on further studies to rule out more alternative behaviour cue explanations. Another good guideline is comparing the studies in infants with chimpanzees. If chimpanzees showed a similar capacity as infants in the same setup, it should be considered as positive evidence for supporting the claim that infants and chimpanzees share a similar ability.

Unlike the two worries I have discussed before, the worry involved in Type three account is not about whether chimpanzees can attribute metal states. People agreed chimpanzees can ascribe mental states to others but disagree on what kind of mental states should be considered. For example, the knowledge attribution theory (Phillips & Norby, 2019; Phillips et al., 2021), which I have presented in chapter 3, claims that chimpanzees can only attribute knowledge states but not beliefs. To settle this disagreement, we need to consider what should count as the mental states that need to be attributed to count as having a ToM, and to what degree the chimpanzees should be thought capable of mindreading. I think mental states involving in ToM at least include intentions, knowledge, belief, emotion, and desire. A full-blown theory of mind at least includes

three levels. The first level is about the ability to attribute goals and intentions to others. This ability enables the agent to ascribe an inner state to others. To think of others' actions as goal-directed means to understand why other individuals performed such actions. It is the foundation for understanding the more complex mental states of others. The second level is about attributing knowledge and perceptions. This ability enables individuals to track what and even how others perceive and know about their surroundings. It will give them the ability to better communicate with each other. The third level is about attributing beliefs which include false beliefs. With the ability to attribute beliefs, chimpanzees will have an extra advantage in predicting others' behaviour, which is the key to why chimpanzees can perform so flexibly in experimental setups. In the following sections, I will examine how chimpanzees performed on each level.

2. Goals and intentions attribution

Understanding others' goals and intentions is essential to engaging in social activities. For chimpanzees, attributing goals and intentions to others will help them understand and then predict what their conspecifics will do. Most researchers would agree that chimpanzees can track the goals or intentions of others since this ability is the fundamental to animals' social life. But they disagree on what should count as a "goal" and an "intention".

In this section, I will separate the discussion of goals and intentions into two parts. In the first part, I will focus on the basic understanding of others' goals and intentions. This involves distinguishing intentional actions from unintentional ones. In the second part, I want to go a step further than simple goals. I will talk about more complex goal attributions in communication and hunting.

2.1 Basic Attribution

The basic attribution of goals and intentions should enable chimpanzees to recognise whether an action is intentional or not. A good starting point is looking into human studies. In human infants' studies, different methods have been adopted in experiments. Mainly, there are three types of methods to check whether they can track the goals and intentions of others. Here, I only give you a basic idea about the following experiments and later I will present them in detail when I compare them with chimpanzees' studies. The first method involves imitation. Gergely et al. (2002) found

14-month-old infants will only imitate the action based on their understanding of the goal of the actor. The second method concerns whether they can distinguish between accidental and intentional behaviour. Carpenter et al. (1998) and Call & Tomasello (1998) used different protocols to show that children from 14-36 months old can recognise intentional as opposed to accidental actions. The third method is about distinguishing between being unable and being unwilling. Call et al. (2004) and Behne et al. (2005) showed children from 9 months old preferred to beg from the agent who was unable instead of unwilling to give them help. If chimpanzees can achieve what human children do in similar setups, there is a very good chance that chimps are using a similar mechanism to do that. The reasoning is based on the idea that if these studies show what they say they do, then infants and chimps use the same mechanisms to perform these tasks. Of course, such an analogy between chimpanzees and human infants is based on the premise that infants have a ToM that explains these findings. To the extent that chimpanzees perform similarly on similar tasks, there is just as good a reason to suppose that they can track goals and intentions as there is to suppose that infants can. And, unless a better competing explanation for these results could be found, we should accept this explanation. In the following paragraphs, I will check how chimpanzees performed on these three types of tasks and how convincing the evidence is.

2.1.1 Imitation Protocol

In the study of human infants, Gergely et al. (2002) put 14-month-old infants into two groups. Half of them watched an adult used her forehead to switch on a lamp because her hands were occupied (Hands-occupied condition). The other half watched the adult use the same method to turn on the lamp while her hands are free (Hands-free condition). Later, the children were given the chance to act on the lamp themselves. The results showed that only 21% of infants in hands occupied condition imitated the actor's unusual head action, comparing with 69% of infants in the hands-free condition. It seems like the infants understand the presenter's intention in both conditions. In the hands-occupied condition, when human presenters used the head to turn on the light when their hands were not free, the infants were less likely to imitate this unusual way to turn on the light. The infants understand the experimenter intentions as being to turn on the light but not as being to use the head to touch the switch. In the hands-free condition, when human experimenters still use this unusual way to turn on the light with their hands free, infants interpreted such behaviour as involving the intention to use their head to turn on the light (since their hands were free and they

could have used them, but used their head instead). Based on how infants will imitate others' behaviour, this shows how infants understand others' intentions.

Buttelmann and colleagues (2007) applied the same imitation protocol with chimpanzees. In their experiments, 12 chimpanzees were put into two conditions: the *Hands Occupied* condition and the *Hands Free* condition like the human experiments. Unlike the human experiment, chimpanzees watched three different tasks instead of just one (this highlights the point more strongly but makes no essential difference to the logic of the experiment). I will only present one of the tasks in detail. In their so-called Head task, they almost copied the same scenario from Gergely. In the *Hands Occupied* condition, the human experimenter used his hands to hold a blanket around his shoulders. In the *Hands Free* condition, the experimenter will drag the blanket on their shoulder and then put his hands on the desk. In both conditions, the experimenter will use his forehead to touch an apparatus so that the light or a sound was produced from it. After the three presentations, the chimpanzees had the chance to play with the apparatus. Similar results were found in chimpanzees. Chimpanzees were more likely to imitate the experimenters' unusual way when the experimenter they had observed turned the light on with their head when their hands were free. It seems like chimpanzees just like infants can understand the intention of human experiments.

This imitation protocol is a good example to show how difficult it can be to provide a simple behaviour cue explanation. I will try to explain the result from the view of a behaviour reading theory. The chimpanzees' behavioural difference depended on the conditions involving at the stage of when the human touch the switch with their head. When the hands of experimenters are hidden, the chimpanzee will prefer to use their hands to touch the switch of the light. In the other condition, when the human's hands are free, chimpanzees will prefer to use their foreheads to touch the switch. Could there be a simple behaviour rule to explain how chimpanzees imitate others in this situation? The simplest strategy is to copy whatever others do. But unfortunately, that is not what happened—in both cases the human touched the switch with their head. The mindreading theory would say that chimpanzees' imitation is based on how the chimpanzees interpret others' intentions. The behaviour reading theory would say something like chimpanzees will imitate the behaviour according to the rules they learned. In this setup, the behaviour rules the chimpanzees might be thought to be following is that when the hands are occupied, use your forehead, otherwise use your

hands. This rule makes sense in general, and it is quite easy for chimpanzees to learn it. But it doesn't fit with either what the human demonstrators do or what the chimpanzees did. Chimpanzees still used their forehead in the hands-free condition but not the hands occupied condition. And the human demonstrators used their heads in both conditions. It is very challenging for chimpanzees to learn what is the rule behind the same behaviour in different situations. It seems that any behaviour rule account should have the chimpanzee behaviour match that of the human demonstrators. But the chimpanzees' behaviour, like the children's behaviour, doesn't match that of the human demonstrators. On the contrary, the mindreading theory enables chimpanzees to use context to learn why others acted as they did, which helps them to understand others' behaviour based on their intentions. So, the ToM based account can readily explain these results while behaviour rules cannot.

2.1.2 Accidental versus Intentional Protocol

In the human study, Carpenter Akhtar, and Tomasello (1998) demonstrated that infants between 14 to 18 months old were more likely to imitate intentional actions than accidental actions. The difference between these two types of actions was marked by a distinctive associated utterance. When the presenter performed an intentional action, s/he will say "there!". On the contrary, in the accidental action performance, the adult will say "Whoops!". Call and Tomasello (1998) used a different method to mark the difference between intentional and accidental actions. The children between 2 and 3 years old learned to use a marker to locate a toy in one of three boxes. In the test, the intentional action was performed by the experimenter deliberately placing the mark on one box. In the accidental condition, the experimenter accidental (by looking away from it) knocked the marker on the box. The results showed the same pattern: children preferred boxes marked intentionally by the experimenters.

In this accident versus intentional protocol, the key is how to display the difference between accidental and intentional. Normally, the difference is shown by different observable cues. For example, in the setup of Carpenter et al., the difference is shown by different vocal expressions ("there" versus "whoops"). Or in the case of Call and Tomasello, the difference is defined by features associated with actions, like whether the marked box was in their line of sight. In the situation of chimpanzees, the behavioural rules theory can readily use these behaviour cues to

explain the result. So, to productively adapt the experiment for chimpanzees, a different sort of method to mark the cases must be used. Wood and his colleagues (2007) came up with a clever plan to try to rule out a simple behaviour cue explanation. In the experiment, a human experimenter will intentionally or accidentally choose one of two containers before him. Then the chimpanzees had the chance to check the containers. In intentional and accidental situations, the experimenter will use his hand to touch the container (on the same side in both situations), at the same time he was also looking directly at it. The difference is presented more subtly. In the intentional condition, the experimenter will reach for and grasped the container in a normal way. In the accidental condition, the experimenter will flop his hand and land the back of his hand on the container. In their result, chimpanzees spent more time with the container that was chosen intentionally than the one that was chosen accidentally. This result will rule out a simple behaviour explanation like chimpanzees will choose the object touched by the agent. Since in both situations the agent touched the container, chimpanzees should not show a difference between these two situations. But the behaviour-reading theory may yet be able to explain this. They could say chimpanzees can distinguish a more subtle difference between behaviours like pointing (the gesture in the intentional situation) and landing (the gesture in the accidental situation). Only pointing means that there is something for them. This finer definition of behaviour rules does make sense. Contacting is a very coarse way to define behaviour. In their everyday life, chimpanzees should be able to use different gestures to associate different meanings. From this experiment, we can see it is tricky to design an experimental scenario to rule out all the behavioural explanations. The nature of control experiments is to set up two or more comparison situations. Unlike medical trials that can make participants unaware of which groups they are in, in animal studies, there always have some visible cues to distinguish different conditions. Behaviour reading accounts always hang on this gap. Sometimes, it works, like in this experiment, but not every time. In the following paragraph, I will present a much tougher case for them.

Buttelmann and his colleague (2012) provided an even better experimental scenario to rule out possible behaviour reading explanations. Their way of doing it is using a "context" to make the difference between accidental and intentional designs. In the experiment, they used the same behaviour movement from human experimenters in the test phase. The only difference came from the previous "context" setup. In the context phase, chimpanzees watch experimenters play with

boxes with lips on them, which need some effort to open. After opening the box, the experimenter will give chimpanzees a grape either from the box they just opened (experimental condition) or from their pocket (control condition). This procedure was played several times with every chimpanzee in two days. One day later, in the test phase, the experimenters will try to open a similar box but different from previous ones before the chimpanzees. Their trying movements will last for two minutes. During this period, cameras will record how chimpanzees from different situations will behave. During the test, no reward was provided. They will only watch the experimenter trying to open the box without receiving anything. This means that chimpanzees need to understand what the experimenter was doing if they want to get the reward inside the box. The results showed that chimpanzees from experimental conditions (where the grape had come from inside the box) behaved systematically different from those the control conditions (where the grape had come from the experimenter's pocket). In particular, the chimpanzees from the experimental condition showed more concern (which mean looked longer in this experiment) at what the experimenter was doing that those in the control condition. It means chimpanzees used their previous experience in context phases to develop expectations about how the experimenter will behave. If chimpanzees have the experience of getting food from experimenters after he or she opens the box, they will think the purpose of opening the box is to give them the grape. Otherwise, they were just lucky to get the food.

This study took a very different approach to settle the worry of behaviour reading. It introduced a context phase before the test trials. In the previous study, the trouble comes from how to present a test without a behaviour interpretation during the test. Buttelmann and colleagues fixed this problem by presenting the exact same movements during tests trial between different situations. During the test phase, chimpanzees in different conditions watched the exact same action from the experimenter. From the perspective of behaviour rules, the chimpanzees should react the same in both conditions. The argument is the following: if chimpanzees share the same behaviour rules, when they were exposed to the same behavioural cue, they should respond similarly. One option for the behaviour reading account in responding to this study is to reject the premise of my argument: chimpanzees from different conditions had different behaviour rules. This is because chimpanzees can learn different behaviour rules from different experiences. During the context phrase, chimpanzees from different conditions did have different experiences: for the experimental

condition group (intentional), chimpanzees got their reward from the box the experimenter opened, while chimpanzees in the control condition (accidental) received their reward from the experimenter's pocket (since the experimenter failed to open the box). Then the question is how this difference between these two conditions will affect behaviour rules. I cannot see a clear answer here. Since both groups of chimpanzees got the reward after the experimenter performed the same action, no matter whether the box opened or not, I cannot see why the chimpanzees should treat the experimenter's action differently.

2.1.3 Unable versus Unwilling Protocol

Another way to recognise intentional action is based on distinguishing the difference between unable and unwilling situations. In the human studies, infants from 9-, 12-, 18-month-old groups showed more patience when the adult was unable to give them the toy than when she was unwilling to do so.

This protocol was first introduced by Call and colleagues in (2004) on chimpanzees. Later the same protocol was repeated to test different species like capuchin monkeys (Phillips et al., 2009), and macaques (Canteloup & Meunier, 2017). In the study by Call and his colleagues, the basic scenario was that chimpanzees were separated from a human experimenter by a glass wall with holes on it. In the beginning, the experimenter will give food to chimpanzees through a hole. Then in the test trials, the experimenter was either unwillingly or unable to give the food to chimpanzees. Cameras will record how chimpanzees responded in different conditions. The unwilling actions were performed by experimenters like refusing to give the food to chimpanzees or just eating the food. For example, when the experimenter refused to give the grape, he will put the grape on a table (making sure chimpanzees can see it) and just stare at the chimpanzee. The unable actions were also presented similarly. For example, the experimenter tried to push the grape through a small hole but failed. During the test trials, there were two measurements to tell the attitude of chimpanzees: how long they will stay in the room and how frequently they will approach the food. In the results, chimpanzees begged more and waited for less time in unwilling conditions than unable ones. These results fit with the assumption that chimpanzees can understand other's intentions. If chimpanzees recognised others' as being unwilling to give the food, they should produce more behaviour to persuade the experimenter, and when that did not work, they should

leave. On the contrary, if the experimenter was unable to pass the food, more begging was not helpful, but waiting longer was desirable. The underling logic is that if someone is unable to give you something, begging him or her would not change the physical reality. But you can wait for a bit to see whether such physical limitations can be overcome later.

Even if the superficial behaviour cues like the direct line of the experimenter's sight were controlled in both situations, there is potentially still some leeway for alternative theories. One possibility for this was investigated in the original paper: whether chimpanzees simply responded differently towards different types of actions without really understand what those actions intended. To address this, the researchers included a control condition without food involved. The experimenters did the same bodily movements in both unwilling and unable conditions but without the food. The results showed that chimpanzees left earlier in no-food condition than the food condition in both the unwilling and unable cases. It seems like chimpanzees were even less interested in the no-food condition. So, the food did make a difference in chimpanzees' perspective and chimpanzees were tracking the intentions behind experimenter' movements. Of course, this doesn't rule out other possible behavioural rules. Nonetheless it provides further evidence that fits well with the intention and goal attribution account, even if it doesn't completely rule out the possibility of behavioural rules.

Summary

In all these experiments, chimpanzees showed a broad and flexible ability to understand the human experimenter's intentions in different situations. Most of them controlled for the worries from the most plausible explanation in terms of behavioural rules. Behaviour reading theories must think really hard to come up with an explanation for each case. Even that is doable, they still cannot pull out a united theory like mindreading to explain all the cases I have discussed before. What about other alternative theories like the minimal theory, can they do better? The minimal theory argued that goals can be recognised from the bodily movement without understanding the intention. They claimed "the units of goal-directed action are events comprising mere bodily movements" (Butterfill & Apperly, 2013). In other words, from these bodily movements, agents can recognise the goals. In their account, the word "goal" is not a matter of intentions. It is "simply an outcome

to which an action is directed" (Butterfill & Apperly, 2013). They did not take any context into consideration. This way of defining goals can only apply to very basic behaviour. For example, when chimpanzees reach a distant object, the goal of this reaching action is the object. But as I presented previous, the same action can be considered differently based on the context. Chimpanzees clearly showed such ability. So, the definition of goal-directed behaviour in the minimal theory of mind is not sufficient to cover the intentional behaviour from chimpanzees either.

1.2 Advanced Attribution

In this part, I will focus on chimpanzees' performance in social contexts. Unlike the previous section, which mainly demonstrates how chimpanzees understand human experimenter's goals and intentions in observation, this section focuses on how chimpanzees understand each other. The previous section focused on a narrow definition of intentions, which mainly test the intentionality of chimpanzees but not the content of their intentions. This is because we don't expect chimpanzees to understand much detail about their human partners' thoughts. But when it comes to the communication between chimpanzees themselves, there is much more going on. In the following paragraphs, I will talk about several situations in chimpanzees' social life and go through one experiment in detail. I will argue that chimpanzees can understand what their counterparts intended to do in all these situations.

Prosocial contexts

The first type of contexts I will consider is in prosocial contexts. The basic idea is whether chimpanzees will help experimenters or their conspecifics to get what they want. This kind of cooperation behaviour is widely observed in human children. Helping others shows not only a willingness to help but also an understanding of others' goals. There is mounting evidence showing that chimpanzees will help conspecifics get food and non-food items in experimental setups. (Warneken & Tomasello, 2006), (Warneken et al., 2007), (Melis et al., 2011), (Yamamoto et al., 2012).

From the current evidence, the helping behaviour can be classified into two categories. The first one is helping others based on their specific request. This has been widely investigated in human children and non-human primates. Here are some examples of such cases: to help others get items out of their reach; to help others move a physical obstacle in their way. Among them, the out-of-reach scenario is well studied in chimpanzees. In those studies, one recipient chimpanzee tried to reach an item out of their range, while the subject chimpanzee close to them can get it. Researchers will record whether the subject will help the recipients to grab the item under their request. Helping behaviour is defined by the following: the other chimpanzee tried to reach towards the item themselves but failed. On the contrary, if the chimpanzee only looks at the item without reaching toward the item, such behaviour is understood as not helping. The results showed that the subject chimpanzees were more likely to provide help in request conditions than no-request conditions. It suggested that chimpanzees offered help because they understood other's intentions. One more additional point that favoured mindreading theory is that in these experiments, chimpanzees were not rewarded for their helping. Normally, to learn a behaviour rule, chimpanzees will receive positive feedback, commonly food. It would be more difficult to explain why chimpanzees performed such behaviour without any obvious benefit. However, behaviour reading accounts can still offer an explanation based on the distinct requesting behaviour. For example, one possible behaviour rule like "get the object to whoever is reaching toward it and others will do the same for you". Chimpanzees can learn such reciprocal principles by observation or from personal experience. That is why the second scenario is important.

The second one is helping others without their specific indications. This is a much more demanding scenario compared to the first one. Without the pointing from the conspecifics, chimpanzees need to figure out what others need by their understanding of the situation. It also reduces the possibility of only using behaviour cues to explain the reaction of chimpanzees. This scenario is the one I want to focus on in this part. Yamamoto and colleagues (2012) designed a brilliant experiment to investigate how chimpanzees understood others' goals. The basic idea was whether a subject chimpanzee can give the right tool to a conspecific who needs it to get the food. In the study, one chimpanzee was put in a predicament that either needed a straw or a stick to get juice. Near this chimpanzee was the subject chimpanzee who was in a different booth from the other chimpanzees. The subject chimpanzee was offered a box that contained seven different tools (includes a straw

and a stick) or non-tool objects (such as a brush). The subject chimpanzee can give objects to the neighbouring chimpanzee through a hole on the wall between them. The subject chimpanzees went through two conditions: *can see* and *cannot see*. In *can see* the condition, the subject chimpanzee can see the situation of the other chimpanzee through the transparent wall. In the *cannot see* condition, the subject was visually blocked by an opaque occluder. During both conditions, the chimpanzee faced with the problem will stretch out their arms through the hole to the subject chimpanzee. The researcher recorded the response of the subject. If the subject chimpanzees understand their counterparts, they should give the right tool (either a straw or a stick) in the *can see* condition but not pick the right one in the *cannot see* condition. The result matched this prediction. According to their first pick, the subject chimpanzees overwhelmingly chose the right tool in *can see* conditions, and they randomly picked an object in *cannot see* conditions. This case is very hard to swallow on the behaviour reading account. Unlike specific requesting cases I presented earlier, the requesting chimpanzees, in this case, used the same body language for help in both conditions. The only significant difference was whether the subject chimpanzee can or cannot see the situation of one needed help.

3. Perception and knowledge attribution

Tracking others' perceptions and knowledge is important part of animals need to do to communicate successfully and more importantly to take advantage of others in the competition. The ability to track others' goals and intentions is useful but not enough if they want to take an extra advantage in competition. For example, in a food foraging situation, when chimpanzees can track where their competitors perceive food being and where they do not perceive food as being, they can nick the unnoticed food, which their competitors do not perceive. As we will see it should not be controversial that chimpanzees can track others' perception or knowledge in this way (for further evidence beyond what we will consider below, see Call & Tomasello, 2008; Krupenye & Call, 2019).

3.1 Perception attribution

The ability of perception attribution relates to chimpanzee's ability to track what and even how other agents perceive their surroundings. One way to classify such an ability is based on visual

perspective-taking. Flavell and colleagues (Flavell et al., 1978; Flavell et al., 1981) distinguished two progressive levels in children, level-1 and level-2 perspective-taking, to describe how children develop their ability to understand how others perceive the world. Level-1 visual perspective-taking is the ability to judge what others can perceive from their point of view. This covers situations like following the gaze of others, to judge whether someone else can see a danger. Meanwhile, level-2 perspective-taking concerns the higher-level understanding of perspectives. Flavell et al. (1981, p. 1) defined it as "an object simultaneously visible to both the self and the other person may nonetheless give rise to different visual impressions or experiences in the two if their viewing circumstances differ". Another way to think about level-2 perspective-taking is by comparing with false belief tests. Instead of thinking of someone holding a false belief, the agent thinks s/he perceives a "false" or different version of the situation before them. As I have talked about level-2 perspective-taking in chapter 3, I will focus on the cases of level-1 perspective-taking in the following.

3.1.1 Gaze Following

To follow the gaze of others is the foundation of perspective-taking. In primates, the gaze is defined by the direction of the head and eyes. In previous studies (Call et al., 1998), by following the gaze of a human experimenter, chimpanzees have been shown to be able to /track the location which is above and behind themselves. A similar result was also found in following conspecifics' gaze (Tomasello et al., 1998). This should be no surprise for both behaviour reading and mindreading. One simple behaviour rule could explain it: directly go for what or where is in the direction of the other's gaze. It is a simple but powerful rule to explain chimpanzees' behaviour. In many situations, chimpanzees did just as the rule predicted. But not all situations can be explained by the rule. A trickier question is to decide how and when to follow others' gaze. In the same paper from Call and colleagues (1998), they found that chimpanzees would only use human's gaze direction in the case when human experimenters know where the food is. In the task, experimenters can see where the food was when it was place into one of two tubes, but they can't see the food when it was put between two bowls. The chimpanzees were always unable to see the food directly from their perspective, but they can see the gaze of the experimenters. The results showed that chimpanzees used human's gaze to choose the tube containing food. But they didn't follow human's gaze to choose between the bowls. This means that they are not following the simple behaviour rule I

mentioned earlier, of just following gaze and going for what is in the line of sight in all circumstances.

Another study (Bräuer et al., 2005) also argued against the idea of behaviour reading. Chimpanzees can follow the gaze to a location behind different barriers. For example, when a human experimenter looked at a window covered with paper and wood which they cannot see through, the chimpanzees preferred to check the place behind the window. It showed that chimpanzees didn't just take the direction of gaze at face value. What interested them is the thing perceived by others, not the gaze.

3.1.2 Food competition

The food competition protocol is one of the best-known experimental protocols in the studies of chimpanzees. Scholars like Hare, Call and Tomasello (Hare et al., 2000), suggested that chimpanzees normally competed for food rather than cooperating to get the food, and so tests of ToM in chimpanzees should employ a food competition scenario. In the food competition scenario, one dominant and one subordinate chimpanzee compete for food. Two pieces of food are placed between them, and one piece of food can only be perceived from the subordinate's view. The subordinate chimpanzees prefer to retrieve the food that can be seen by them but not by the dominant individuals. This preference of subordinate chimpanzees has been repeatedly in different studies (Hare et al., 2000), (Hare et al., 2001), (Bräuer et al., 2007), (Santos et al., 2006). Behaviour reading theorists came up with an alternative, behavioural rule explanation, however. They suggested that a rule like the following could explain the behaviour: whenever the food was under the gaze of the dominant chimpanzees, the food was "contaminated", so the subordinate chimpanzees will avoid such food. The rule did explain why subordinate chimpanzees preferred the food that wasn't seen by dominant chimpanzees. However, Hare and the colleague (2001) made an updated version to test such possibilities. In this new setup, in which a dominant chimpanzee who saw the baiting of the food was replaced by a new dominant. Notice that this should not change the prediction of the behaviour rule theory, since the chimpanzee that is replaced will have already "contaminated" the food they saw hidden. But the ToM account will now predict that the subordinate chimpanzee should be more willing to approach the food, since they believe that the "new" dominant chimpanzee doesn't know about the hidden food, and so the subordinate

chimpanzee should be able to get more food in this situation. This is exactly what happened: the subordinate chimpanzees retrieved a larger percentage of food when they competed with the naive dominant who had not seen where the food was placed. This means that not only the gaze of dominant chimpanzees affected the choice of subordinates but also the identity of the dominant (and what they have seen). This is a problem for the behaviour reading explanation above. Because they only focused on the direction of gaze without concerning whose gaze was. To patch the theory, the behaviour rule will at least need to include reference to whose gaze should be counted. In this case, only the gaze from the dominant chimpanzee on the field should have the ability to "contaminate" food.

Chimpanzees have shown an even deeper understanding of others' perception in cases (Melis et al., 2006; Hare et al., 2006) where they can cheat others based on their perspective. Hare et al. (2006) showed that chimpanzees would choose to get the food through a route hidden from their human competitors. Another study, by Melis and colleagues (2006), provided an even more astonishing form of deception behaviour from chimpanzees. This experiment different in several ways from Hare et al. study above, but the most significant difference is that, in Hare's case, chimpanzees used visual information to deceive others, in Melis' case, chimpanzees used acoustic information to achieve that. I see Melis' case as revealing a more interesting aspect of chimpanzee's ability of perspective-taking. I will mainly focus on this study. In their setup, chimpanzees competed for food with a human experimenter who sat inside a booth. Food was placed on each side of the experimenter. Chimpanzees could reach the food through two separate tunnels. Each tunnel leads to one piece of food. If the human experimenter found the chimpanzee trying to get the food through tunnels, she would take away both pieces of food. In the test, the experimenter put her head down, giving chimpanzees the chance to steal the food through the tunnel. However, one of the tunnels made a loud noise when the chimpanzees opened the flap on it, and the other tunnel remained silent when opening it. The noisy tunnel was visually different from the quiet one. It turned out that chimpanzees tended to avoid the noisy tunnel when a human sat before them but not when nobody was there. Such result suggested that chimpanzees can use information whether a competitor can or cannot hear to avoid being detected. The most obvious behavioural cue is that making noise will lead to a loss of food. Chimpanzees can learn this rule by familiar processes. When they made noise by reaching through the noisy tunnel, the

experimenter took the food away. However, the researchers designed an additional experiment to rule out this type of behavioural cue explaining the chimpanzees' behaviour. In the experiment, the only difference was where the noise came from. In this setup, the noise came from a cell phone after the chimpanzees get the food through one of the tunnels. If the chimpanzees behave according to the behaviour cue, they should avoid the tunnel which will cause the noise. But the result showed that chimpanzees did not have a preference between the noisy and silent tunnels. This additional experiment excludes the hypothesis that noise will lead to loss of food. It showed that it was not only that there was noise that mattered but also where the sound came from. This new setup may exclude one behaviour rule hypothesis, but it certainly cannot rule out them all. For example, chimpanzees may learn that the source of the sound is also mattered to get the food. It is certainly possible, but the behaviour rule supporters need to provide more cases following such rules.

3.2 Knowledge attribution

Knowledge attribution is deeply connected to perception attribution. In many situations, what animals know is largely based on what they perceive. In this part, I will talk about situations which are more complex than the previous cases. Perspective-taking is only one of many ways to ascribe knowledge to others. This is especially true in communication. You can come to someone know something through many other ways. For example, when you excitedly ask your friends to look at the beautiful sunset, you can tell that your friends have understood you not only by the fact they are looking at the right direction but also in other ways like if you see that they take out their phones to take a picture of it or you hear them shouting joyfully. So, the attribution of knowledge can be done by tracking other signals instead of something like following gaze. In this section, I want to present a case how chimpanzees attribute knowledge to their conspecifics in communication.

Alarm Signal

In the wild, chimpanzees use specific vocal calls or body signals to inform other members of the coming danger like snakes. Such alarm calls are found in wild chimpanzees. Normally, chimpanzees are more often to give alarm signals when they are with kin conspecifics. Moreover, several studies have shown that chimpanzees make alarm signals not based simply on the presence

of a predator but rather on whether others have perceived the danger and how they have reacted (Crockford et al., 2012; Schel et al., 2013; Crockford et al., 2017; Graham et al., 2020). I want to focus on one such case in the following.

Crockford et al. (2017) put a snake model in the anticipated travel path of chimpanzees to induce an alarm call scenario. In the wild, chimpanzees make an alarm signal to notify others when they encounter a snake. The question is what would affect chimpanzees' alarm signals. This study was designed to test whether chimpanzees modulated their alarm signalling based on what their conspecifics know about the danger. In the setup, a snake model was placed behind a log on the regular travel path of a group of wild chimpanzees. The researchers labelled the chimpanzees' alarm signal used in this case, *marking*. They defined marking as involving cases where the subject chimpanzee (the signaller) repositions itself to have direct visual access to both the danger and the receiving chimpanzee (the receiver). The receiver should also have seen the snake. Then the signaller should turn their head between the danger and the receiver without doing any other things. This definition of marking makes sure that the subject chimpanzee is truly signalling alarm instead of producing other general communications. The experimenters then recorded how the signallers and receivers behave. They proposed four competing hypotheses to predict what may happen in this situation. Here, I want to focus on the contrast between the mindreading theory versus the behaviour reading theory, which involves three of the hypotheses. The knowledge hypothesis is a component of the mindreading theory. If the signaller can attribute knowledge to the receiver, the signaller should inform the ignorant rather than the knowledgeable others. So, this mindreading hypothesis should predict that the signaller will produce an alarm (the marking behaviour) for the receiver when the receiver doesn't see the snake and the signaller should stop marking once the receiver sees it even if the receiver keeps approaching to the snake. The other two hypotheses are the signaller habituation hypothesis and the receiver behaviour hypothesis. The habituation hypothesis is based on the habituation effect which says that the chimpanzees will be more active (make more alarm signals such as marking or calling) when they are first exposed to the snake and become less active with more exposure. Based this hypothesis, the signaller should mark based on his own experience with the snake and marking behaviour should be unrelated to how the receiver behaves. The receiver behaviour hypothesis is that the signaller gives alarm signals by monitoring the receiver's behaviour, such as approaching the snake. It predicts that the signaller will stop

marking once the receivers are moving away from the risk (in this case, the snake) and that they should continuing to mark while the receiver continues approaching. The results from observations of how chimpanzees actually behaved in this experiment only matched the mindreading hypothesis. The signaller chimpanzees were more likely to stop marking once the receiver knew the existence of the snake even if the receiver kept approaching it.

4. Tracking false beliefs

The false belief test has long been considered as the golden standard of ToM. The magic of false belief tests relies on the causal role those false beliefs played in the reasoning process. The original idea of false belief test came from Daniel Dennett (1978). If an agent holds a false belief, he or she will predictably act inappropriately in certain circumstances. For example, in the Anne-Sally test, Anne, if she can attribute false beliefs, will expect Sally to search in the wrong place. People have no issues with the classical verbal version of such test. But when it comes to non-verbal versions, all sort of different disagreements shows up. This is because non-verbal versions are based on behavioural cues to tell what the subjects expect. For example, the minimal theory of mind (which was discussed in chapter 2) suggested that to pass a false belief test, the subject doesn't need to represent the other's false belief. There is shortcut way to track false belief by instead representing registrations. Also, as I suggested in chapter 1, non-verbal false belief tests are not immune from alternative explanations based on behaviour cues. These sorts of problems become worse in animal studies because non-verbal tests are the only versions of false belief tests available in this case. So, false belief tests are not as clean cut in the case of animals as they are in human research. Nevertheless, it is still a very high standard experiment protocol which has demonstrated its important in human studies.

Back to the study of false belief in animals, only recently, evidence has started to support the claim that chimpanzees or even monkeys can track false beliefs (Krupenye et al., 2016; Buttelmann et al., 2017; Kano et al., 2019; Hayashi et al., 2020). Earlier studies had only yielded negative results (Call & Tomasello, 1999; Kaminski et al., 2008; Krachun et al., 2009; Krachun et al., 2010; Marticorena et al., 2011). The contradictory results between recent and previous studies make the

discussion about the false belief attribution inconclusive. In this section, I will first introduce the recent works on false belief tasks in chimpanzees, and then talk about how they can shape the topic.

As early as 1999, Call and Tomasello had tried to test chimpanzees with the same nonverbal false belief task which was used to test human infants. But they only found human infants were capable of false belief attribution. Later, Krachun and her colleague (2009) organised a competitive false belief test for both human children and apes. It still produced negative results in apes but not humans. One of the first positive studies about false belief attribution came out in 2016. Krupenye and colleagues (2016) designed two false belief scenarios and used anticipatory looking (AL) as the indicator of chimpanzees' prediction in both cases. Later, Buttelmann et al. (2017) adopted a different arrangement to show chimpanzees understanding others' false beliefs by using active behavioural measures. Kano et al. (2019) adopted a very different protocol to test false belief attribution which I have already presented in chapter 1. Hayashi and his colleagues (2020) using a similar experimental design as Krupenye et al. showed that macaque monkeys can also attribute false beliefs. In the following, I want to present the study from Krupenye et al. (2016), because it is one of the first positive cases and many later experiments followed the same protocol introduced in this study.

Three species of apes (chimpanzees, bonobos, orangutans) were included in two experiments. Both experiments involved the same process and used the same anticipatory looking measure. So, I will only present the first experiment. Before the test, apes went through a familiarisation phase. They watched a short movie which introduces the scenario that a human agent (actor) competed with an ape-like character, which the experimenters referred to as King Kong (KK). The actor placed a goal object in one of two locations, then KK robbed the object and hid it in the other location. After KK left, the actor retrieved the object again. This competition for the object is repeated again in the movie to make sure the subject apes know the actor want to get the target object. Later in the false belief test, subjects will watch one of two movies introduced a false belief scenario. In one of movies, after KK robbed the object from the actor, KK will hide it in one location then move it to the other location, where all of this happens while the actor is present and watching. Then KK will chase the actor out of the room. After the actor left, KK will take the object away with him. At the end, the actor returned to search for the object and the movie ends at the frame

where the actor's hand reaches out in the central between two locations. The first look of subjects was recorded as the indicator where they expect the actor will search between two locations. If subjects think the actor held a false belief, they should expect the actor reach for the location where he had witnessed KK put in the object. Based on the anticipatory looking method, the subjects should first look at the target location (which is where the actor falsely believes the object should be). The results from both false belief tests showed that all three types of apes are more likely to look first at the target location as the mindreading theory suggested.

Like the previous studies, the experimenters designed their setup to rule out a simple behaviour cue explanation such as that the actor searches for the first or last location where KK hid the object. In the false belief test I have introduced above, the last location is the target location where the object was put into. But in the second false belief test they ran, they change the proceeding to make the first location is the target by putting the object in the first location. The subject apes in both tests anticipated that the actor would search for the object in the target location where he falsely believed the object should be, and such results is contradicted with either followed the first or last location. This result also argues against the knowledge attribution theory. KAT, as I presented in chapter 3, claims that the subject cannot tell the difference between others holding a false belief and being in a state of ignorance. According to KAT, the subject predicts the agent who hold a false belief will behave like he is ignorant about the location. So, KAT predict the subject attribute ignorance to the actor, which means the subject apes should look randomly between the two locations. This prediction from KAT is also contradicted by the results.

But as you may guess now, some people are still unsatisfied. For example, Horschler et al. (2020) question the validity of the anticipant looking method (AL) adopted in this study. They argue that such a measure is less reliable than the measure involved in something like expectancy violation. The main problem of AL is that there isn't a consistent way to measure the first look. Different experiments used different ways to measure the first look. For example, among the false belief test in chimpanzees, Krupenye et al. (2016) used a 4.5s response time-window to measure the first look. However, Kano et al. (2019) used a 6s time-window based on the same design. These concerns about the methodology are reasonable ones which I have to admit only more data can

answer. But at least, such positive evidence revives the possibility that apes may attribute false belief to others which many theories cannot explain.

Summary

In this chapter, I have provided evidence to support the claim that chimpanzees can attribute mental states including intentions, perceptions, and beliefs to others. Unlike some researchers in the field who only focus on false belief tests in chimpanzees, I have broadened the scope of attributions of other mental states such as goals and intentions that I am interested in. This is because, even if the false belief test is the golden standard in mindreading studies, other alternative theories can still claim chimpanzees pass such a test but deny they possess a ToM. As I have discussed in chapter 1, the behaviour reading account can comfortably explain such evidence using behaviour rules. The minimal theory of mind account also has no problem with chimpanzees passing false belief tests. Instead, the minimal theorists actually take the advantage of the importance of the false belief test for many researchers in these debates in order to design their theory in a specific way to satisfy such tests. The only theory that cannot accommodate the false belief test is the knowledge attribution theory, which ironically is based on accepting a subset of the elements of more standard mental attribution (ToM) theories. So, as I have emphasised in this chapter, it is important to examine other type of evidence bearing on mindreading abilities in order to grasp the whole picture in the case of chimpanzees. Experiments involving attributions of intentions and perceptions provide some very convincing evidence to support the standard mindreading account. All of the other alternative theories that I have discussed in this thesis lack the explanatory power that this theory has, when considered in light of this broader range of evidence. Based on the current evidence, I conclude that the mindreading theory is most convincing theory for explain chimpanzees' behaviour.

Chapter 5 ToM in Corvids

As I noted in the introduction to the thesis, corvids are a songbird family that includes crows, ravens, and jays. What makes them so interesting in mindreading studies is their high intelligence and rich social lives. One of the biological indicators for intelligence in animals is the relative size of brains compared to their body size. Crows have the similarly big brains in this relative sense as chimpanzees and an even larger forebrain (Emery & Clayton, 2004). Putting the size of their brains aside, corvids also show extraordinary skills in tool use (Rutz et al., 2016; Bird & Emery, 2009b). The ability to use or even make tools is regarded as a sign of high intelligence among animals. Besides tool use, corvids matched or even outperformed chimpanzees in many other cognitive tasks such as planning (Raby et al., 2007; Kabadayi & Osvath, 2017) and understanding certain types of physical causation (Bird & Emery, 2009a; Jacobs et al., 2015).

In this chapter, I will first talk about why corvids are so interesting in the study of ToM. Then I will present an overview of studies of how corvids performed in different mindreading tasks. I argue that even if the evidence supporting the mindreading theory in corvids is not as strong as it is in chimpanzees', it is the most persuasive theory to explain their behaviour. Last, I will compare chimpanzees and corvids on mindreading to see how mindreading can be achieved in different species.

1. Why corvids

There are mainly three reasons making corvids an interesting case in the studies of mindreading. First, they are highly intelligent animals from many perspectives. This gives them the ability to act flexibly and skilfully in different contexts. Second, they perform many different cognitive skills like planning and episodic memory. This unlocks their potential in other social cognitive abilities. These are reasons to focus on corvids in particular. The third reason isn't particular to corvids, but would apply to any other type of animal that is less closely related to us than other primates—this reason is that, if corvids are able to attribute minds to others, it will open another new door to thinking about mindreading in convergent evolution. It will help us better understand how mindreading evolved and how it works with different biological bases.

1.1 High intelligence

Ravens have long been rumoured to be intelligent animals, going back as far as ancient fables. In Aesop's fable, a thirsty raven managed to drink water that was out of reach by dropping pebbles to raise the level of the water. It turns out that this tale is not very far from reality, in a study (Bird & Emery, 2009a), researchers found that rooks can use stones to raise the water level which led to their being able to reach a floating worm. It is not surprising that people would think of such behaviour as intelligent, because tool use was historically believed to be a uniquely human characteristic (e.g., Oakley, 1964). More recently, many kinds of tool use or similar behaviours have been found in primates (e.g., Boesch & Boesch, 1990), birds (Hunt, 1996), even reptiles (Dinets et al., 2015) and fish (Brown, 2012). But the most proficient and intentional tool using behaviours are still very rare among animals. Corvids are believed to be one of them. Their sophisticated tool using skills have been studied by many researchers both in the wild and in labs. For example, in the wild, New Caledonian crows were found to manufacture hooks from leafy twigs chosen by themselves (Hunt & Gray, 2004). In the lab, rooks showed a flexible ability to choose either stones or sticks to get food from a tube (Bird & Emery, 2009a). All these tasks involving tool use required the subjects to understand the basic physical properties of tools and how to solve the problem at hand by using such properties. These tool-using skills ranged from basic tools selection to more advanced tools manufacture, and even metatool use, which means using one tool to get access to another tool to use to access food. The most interesting case among these is the case of metatool use (Clayton, 2007; Taylor et al., 2007; Gruber et al., 2019). Metatool involves an inverse reasoning process like a detective uses in trying to solve a crime. To get the final reward, you need to first figure out which tool is the right one, then you need to such reasoning again to find out which tool is needed to obtain that tool. I want to show you the case of Gruber et al.(2019)

In their study, the crows were put in a task that required them to get either a stick or a stone at first. Then if they picked the right tool, they could retrieve the reward. For example, if crows were put in the stick condition, they were provided with a short stick initially. Two apparatuses containing different tools were presented. Only one of them had the right tool to access the baited meal in the third apparatus. The initial short stick can only retrieve the tools but not the meat. Crows must use

the short stick to get the stone and then use the stone to get the meat. Most of the subject crows correctly accessed the meat without any error. The significance of this experiment is that crows need to choose the right tool before they actually got it. To make the right choice means that crows need to understand the problem and plan several moves ahead. One reasonable way to achieve such skills is mental representation plus reasoning based on such representations. By representing the tools in mind and figuring out which tools is suitable for which situation, the crows can solve the test in their mind before touching the physical test. To form mental representations and manipulate such representations is bedrock for more complex cognitive skills like ToM.

Reasoning and Causation

Since corvids can flexibly choose or create their own tools to get what they want, it is natural to wonder how they do that. One obvious answer is they understand the causes and the effects involved in their tool use. They know what kind of tools can cause a certain effect to achieve their goals. But many people don't buy this type of causal understanding hypothesis (Penn et al., 2008; Taylor et al., 2014). Penn, Holyoak, and Povinelli (2008) claimed that animals at best could only understand first-order perceptual relations but not the higher-order relations based on the perception relations. The first-order perceptual relations are about how different behavioural cues are associated with different outcomes. The higher-order relations refer to abstract concepts. For example, dogs started to salivate when they heard the bell ring in Pavlov's experiment. This can be explained in terms of first-order relations: dogs associate the bell's ring with the food. This distinction between animal cognition and human cognition, with animal cognition being thought to be limited to first order perceptual relations, was not limited to causal relations but to many other fields such as ToM. Even without causal cognition, animals still could have ToM, but I am interested in digging into causal behaviour because if corvids have causal understanding this means they are not restricted to first order perceptual associations, and that increases the plausibility of the ToM hypothesis. Meanwhile, since most supporters for the behaviour reading theory like Penn and Povinelli reject the idea that non-human animals can understand causation, this would also limit how and what animals can learn in terms of behaviour rules. So, I want to talk about how corvids perform in understanding causation, and the best cases to support such arguments are in tool use. The debate about causal understanding in animals continues as unsolved, but the positive

evidence is accumulating (Jacobs et al., 2015; Taylor et al., 2010; Smirnova et al., 2015; Seed et al., 2011; Tebbich et al., 2007).

1.2 Cognitive skills

Beyond tool using and reasoning, there are other interesting skills corvids have mastered. I want to talk about two other aspects of their cognitive skills: episodic-like memory and future planning. Both abilities are essential to their caching behaviour. In the real world, corvids will store food when it is abundant like in autumn, and later retrieve it in the winter. Perhaps the most impressive example of this involves Clark's Nutcrackers, who can cache as many as 80,000 items among thousands of locations. Clark's Nutcrackers live on these caches through the whole winter and spring. To achieve that, they need to not only remember where these caches are but also plan for how many are needed.

Episodic memory refers to the ability to remember the specific aspects from past events. It helps us recall where and when a particular event happened. The ability had been thought to be unique to humans (Tulving, 1972). However, if we put the complicated issue of the phenomenal character aspect of episodic memory aside, such episodic-like memory has also been found among corvids (for reviews Pause et al., 2013; Miyamoto Gómez, 2021). Corvids' ability to remember certainly will affect how they protect and pilfer caches. Another factor related to caching behaviours concerns planning. Specially planning for the future beyond one's current needs and motivations has also been thought to be a uniquely human ability, but (Suddendorf & Fletcher-Flinn, 1997, 2007) showed that ravens would cache more food if they expected no food tomorrow.

1.3 Convergent evolution of ToM

As Emery and Clayton (2004) proposed in their paper, the similarity between chimpanzees and corvids may provide a case of convergent evolution in some cognitive abilities. They claimed that corvids and chimpanzees share similar performance in the tests related to tool manufacture, mental time travel, and social cognition. Here, I only want to emphasise ToM. I have already presented many cases in previous chapter showed both chimpanzees and corvids passed similar ToM related tests. Based on other similarities mentioned by Emery and Clayton (2004), it is reasonable to

assume both animals may share the same cognitive skills in relation to ToM. This may because both chimpanzees and corvids were under the same social pressure in evolution. For example, both need to understand the relationships between different individuals in a large group. Also, currently only these two species consistently demonstrated their ToM ability. Since birds and primates share a common ancestor 300 million years ago (Burt et al., 1999), it is long before the presence of ToM. If chimpanzees and corvids do share the same sort of mindreading ability, this provides a case of convergent evolution which means both species evolved such an ability separately. A reasonable explanation is convergent evolution of ToM, which means both species evolved ToM separately. Of course, there are a lot of assumptions to digest here. I wouldn't push this idea too hard, but it provides an interesting comparative study opportunity here.

All of the above points give an authentic motive to dive into the question about how corvids did in mindreading tests and what is the most reasonable theory to explain their performance. In the following sections, I will mainly focus on two aspects related to mindreading abilities: perception and knowledge attribution and desire attribution.

2. Perspective taking and knowledge attribution

Since corvids are equipped with similarly rich cognitive capacities as chimpanzees across a range of cognitive domains, it is natural to wonder how corvids perform in mindreading tasks. In this section, I will focus on the elementary aspect of the mindreading ability which is the ability to attribute perception and relative knowledge to others. Most of the studies about corvids mindreading are related to their caching behaviour. Food caching is a very common behaviour among corvids which enable them to store food for future use. Cached food is a major part of their diet and key for their breeding (Bossema, 1979). Since food caching is so important for their living and reproduction, it is no surprise that corvids will use all their intelligence on their caches. Most studies are designed around such caching behaviour. In the following, I will argue that corvids protect their own caches and pilfer others' caches by attributing perception and knowledge states to others, and that this type of explanation is superior to other potential explanations such as accounts given by the behaviour reading theory or the minimal theory of mind account.

2.1 Perspective taking

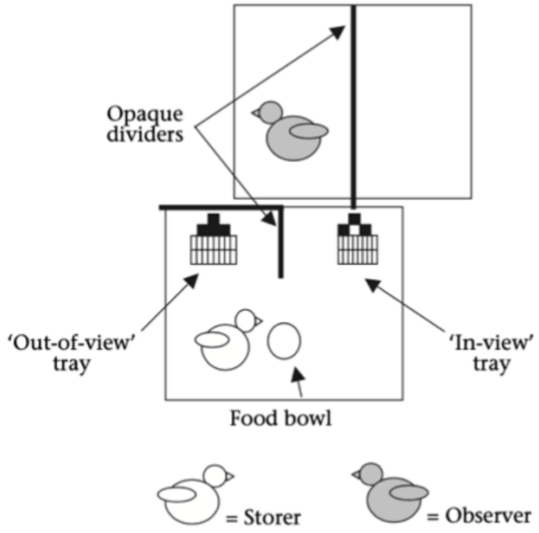
Corvids have been found to take many different strategies to protect their food caches. These cache protection strategies have included eating the food (Emery et al., 2004), hiding food in the shade (Dally et al., 2004), caching food further or out of other's view (Dally et al., 2005), re-caching food in new locations (Dally et al., 2010), and hiding food behind barriers (Legg & Clayton, 2014). Their cache protection behaviour is not limited to being concerned by visual presence of conspecifics. Scrub jay are also, for example, concerned what others might hear, and are concerned about their own sounds (plausibly to not alert conspecifics). For example, scrub jays (Stulp et al., 2009; Shaw & Clayton, 2013) will limit auditory information to conspecifics during caching. And perspective taking is not only shown in cache protection but also in cache pilfering. The pilfering corvids also take others' visual and acoustic perspectives into account when they want to steal others' caches (Shaw & Clayton, 2014; Clary Dawson, 2016). Corvids' protection and pilfering behaviour are ubiquitous, but which theory is best suited to explain them is less clear. To decide that we must dig into some of the details in experiments.

Taking the study of Dally, Emery and Clayton (2005) as an example, the birds in their study preferred to cache food in a place further away from the observer. In their setup, the subject birds were given the chance to cache the food in two trays. One tray was placed near the observer's cage, and another one is placed further from the perspective of the observer. The observer was placed in a separate cage next to the subject's cage. In the observed caching condition, the subject birds need to cache worms under the presence of an observer. In the private caching condition, an opaque barrier was placed between the storer's cage and the observer's cage, which meant that the observer has no visual access to the storer's cage. The subject birds were also given the same trays and food to cache without the presence of an observer. Later, birds in both conditions were given the time to recover their food in private which means no observer in their view. Researchers counted how many worms were in each tray before and after recovery. The results showed that before recovery, birds in the observed condition stored more worms in the far tray than the near tray, but not in the private condition. During recovery, birds in the observed condition also tended to recache more worms from the near tray than the far tray, but not in the private condition. These results were consistent with the mindreading theory. If the caching birds can attribute perception to the observer, being further away from the view of the observer is a more secure place to store the food. This

logic goes for both caching and recaching behaviour in birds. But this single case could also be explained in terms of the behaviour reading theory. One of the possible behavioural rules is an "avoidance hypothesis" which is that the storer will try to get away and stay away from other birds when caching food. Such a hypothesis derives further support from another finding from the experiment. During the caching, storers spent more time on the far side of the cage when an observer was there. Following this behavioural rule, the birds cache more food in the far tray in order to avoid caching near other birds. But this explanation fails to account for the difference during the recovery stage when no observer was present. In the recovery stage, the subject birds tended to relocate more items from the near tray than from the further tray. This finding cannot be explained by the rule that birds try to avoid other birds when caching food. I think it is fairly easy to patch that loophole by expanding the previous behaviour rules: the caching birds will not only try to stay away from other birds but also try to stay away the place was closed to them. On such an outstretched "avoidance hypothesis", caching birds may view the place close to a conspecific as "poisonous" which should be avoided if possible. This is case if the observing bird is no longer present, this rule could be that caching birds avoid caching anywhere near where any other birds have been (recently). This can explain why the caching birds preferred the far tray in both caching and recovery stages. One consideration that might support this sort of hypothesis is that it is not implausible that birds with experience of losing caches might notice that they are more likely to lose them when they had been cached closer to the area where a conspecific had been present.

The researchers from the same study came up with an alternative setup to test the hypothesis that the bird doing the caching would avoid caching in a place nearer to the observer. In this new setup, two trays were placed at the same distance to the observer, and one of the trays was surrounded by an opaque barrier (this tray was labelled the "out-of-view" tray). The other tray was placed in the view of the observer. You can find the details in Figure 3 below.

Figure 3



Note. This picture was produced by Dally et al. in 2005, showing the experimental setup for caching strategies of the storer. From "Cache protection strategies by western scrub-jays, Aphelocoma californica: implications for social cognition," by J.M. Dally, N.J. Emery, and N.S. Clayton, 2005, *Animal Behaviour*, 70(6), p1259 (https://doi.org/10.1016/j.anbehav.2005.02.009)

If the avoidance hypothesis was right, the caching birds should not show any difference between the two trays for caching and recaching when an observer was present. But if the birds can attribute perception to the observer, they should prefer the "out-of-view" tray because only this tray was not perceivable by the observer. The results supported the mindreading theory. The caching birds cache more food in the tray that is out of sight from the observer. This rules out the avoidance hypothesis above. However, I think the behaviour reading theory can adapt their account to explain the results from this new setup. The behaviour reading theory can suggest a new behaviour rule that not only takes account of the distance of conspecifics from potential caching locations but also considers which locations are in the direct line of sight of conspecifics. The direct line of sight of observers was one of the most important behavioural cues in many of conditions that I presented in chapter 1. It is also one of the most useful rules in animal experiments. According to direct line of sight hypothesis, the subject only needs to judge whether the observer has a direct gaze on the item. Such judgment doesn't require the subject to take the perspective of the observer because they can tell what is in the observer's direct line of sight directly from the spatial relations between the observer and the item. Combining the rules in these two hypotheses together, we can get something like this: cache outside of the direct line of sight of others, and if that is not possible,

cache in the place further away from them. With this kind of rule, the caching birds' caching behaviour can be explained without supposing that they are attributing perceptions to others.

But the problem with this approach is that adding more behaviour rules has its limits as I have discussed before. For one thing, there is the challenge of explaining how corvids actually can learn such rules with very limited experience. In the lab, these hand-raised birds did not have much chance to experience caching and pilfering occurring when caches were at different distances from an observer. Some may grumble that the mindreading theory faces the same challenge here, namely how should the corvids behave in different circumstances. Mindreading theory would simply suggest that the subject bird will do whatever it can to reduce the chance of exposing the food to the observer. Because the subjects can take the perspective from the observer, they would know what to do in different situations. The second disadvantage facing the suggestion that we can address the experimental results here by just combining behaviour rules is that there are too many factors affecting the caching behaviour of corvids. First, protection behaviour from caching corvids can be triggered by lots of factors other than proximity of an observer. Things like the light conditions also affected where caching birds decide to cache the food. In the study of Dally et al.(2004), corvids preferred to cache more items in a tray where there was less lighting than one with better lighting when an observer was present. But this preference disappeared when corvids cached in private. Second, corvids also use a variety of methods to protect their caches beyond re-caching. For example, when storers observed some of their caches were stolen, they tended to eat more portions of caches instead of recaching them. Combining these two factors together, the task in front of the behaviour reading is to connect the caching scenes and the desired protection behaviour with each other correctly. More specifically, given that many factors can influence protection behaviour, and protection behaviour can take many forms, to capture the full range of experimental results requires the behaviour reading account to create increasingly complex and implausible behaviour rules. However, if the corvids can take the perspective of the observer, the solution becomes much easier.

What about the minimal theory of mind account? This account may explain the protection behaviour by referring to the encountering and registration mechanism that it posits. But it has its own shortcomings. One of them is that the minimal theory is only built on visual information and

lacks the ability to track other sensory information like audition. There is evidence showing that corvids can track others' acoustical information (Shaw & Clayton, 2013; Shaw & Clayton, 2014; Stulp et al., 2009). In the following, I will present one of the studies to see how the minimal theory suffers in light of such evidence.

Shaw and Clayton (2013) demonstrated that Eurasian jays will act so as to limit the auditory information available to conspecifics during caching and pilfering. In their experiment, there were two types of caching trays. One had a noisy gravel substrate and the other one had a quiet sand substrate. The jays were given the chance to cache the food in three conditions: an alone condition in which no observers were presented, a seen and heard condition in which observers had visual and acoustical information of the caching, a heard-only condition in which pilferers could hear but not see the caching bird's caching. Shaw and Clayton compared jays' caching preferences between these different conditions. The results showed that in all conditions the jays tended to place the food in the quiet substrate. However, such tendencies were more significant in the heard-only condition compared to the other two conditions. This means that jays will cache more portions of food into the quiet tray in the heard-only condition. Between the "alone" and the "seen and heard" condition, there was no such preference. Moreover, a similar preference for the quiet tray was found in pilferers as well in this study. This time the subject is the pilferer instead of the cacher. They were given the chance to observe where the cacher will store the food. The pilferers were then provided with the opportunity to look through two peepholes in an otherwise opaque barrier. When the pilferer stood at one of the peepholes, a bell attached to it would make noise. The other peephole did not have such a bell and so would not make a noise when a pilfer stood before it. The results showed that pilferers preferred to use the quiet peephole rather than the noisy one to observe cachers during their caching process. This result showed that corvids can use auditory information to adjust their caching and pilfering behaviours. This kind of case is not a big problem for the behaviour reading theory in general. Because this theory can explain the results using a simple behaviour rule like when there is noise during the caching or the pilfering, choose somewhere quieter. But the minimal theory of mind account has to make quite a big change to extend its principles to cover the case of auditory information. Their original principles only cover the case of visual information. Their concepts of encountering and registration only refer to visual signals. So a substantive extension of the account would be required to address auditory information.

Moreover, I can imagine other sensory channels like olfaction should also be taken into consideration. The minimal theory could deal with such challenges by adding further principles to cover cases where information from these other types of sensations is at issue. So, the problem here does not necessarily kill the account, but the problem raised by cases involving other types of sensory input does damage the promise of the minimal approach. It must add more principles to just cover the additional sources of sensory information as I explained in more detail in chapter 2. By contrast, the mindreading theory only needs to suppose that corvids assume that other birds share their own perceptual abilities. That is, they will just assume that observers naturally can perceive the same things that they themselves are capable of perceiving (in the observer's circumstances). But the minimal theory has to build in from scratch each such ability to perceive and their limitations, using only minimalist resources analogous to encountering and registration.

2.2 Knowledge attribution

As I discussed in the previous section, corvids are very sensitive to the visual or auditory presence of other conspecifics. Their cache protection behaviour is also sensitive to many other factors. I will classify all these extra factors related to perception as involving "knowledge". For example, Scrub-jays tracked which specific individual was observing them and behaved differently based on bystander's identity (Dally et al., 2006; Bugnyar & Heinrich, 2005; Bugnyar & Heinrich, 2006; Bugnyar, 2011; Shaw & Clayton, 2012). In these experiments, both cachers and pilferers tracked the identity of who was observing them or who was doing the caching. Moreover, they were able to track features associated with the social identity of the cacher or pilferer such as whether they were dominant or not.

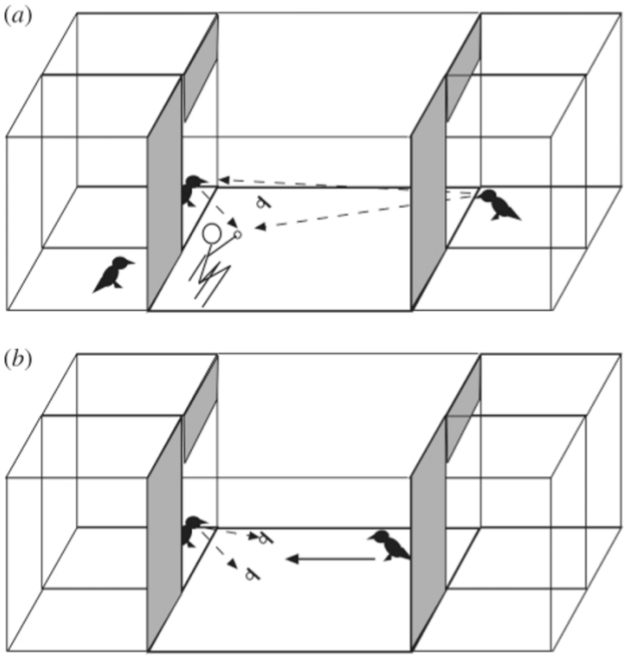
Let's look at a sample study. Dally, Emery, and Clayton (2006) designed a study to test whether Western scrub-jays tracked who was observing them when they cached the food. In their setup, the cachers had to cache the food in one of two trays under the observation of an individual bird A, and later cachers had to cache the food in the other tray when a different individual bird B was observing. After a while, storers recovered the cache in four different situations: in the view of either bird A or bird B, in the private which means no observer present, and in the view of a "stranger" who had not been present in either stage of the caching. The idea behind their study is that if the cacher doesn't care about the identity of the observer, they should behave similarly in

these situations regardless of who is observing them, whether it is bird A, bird B, or a stranger. On the other hand, if they do care about who is observing them, they should adjust their re-caching strategy based on the identity of the bird that is present. Several interesting re-caching strategies were found by counting how many items were re-cached. First, in view of a "stranger", cachers barely re-cached any food from either of the trays. Second, in the presence of either A or B, cachers only recached the items from the tray where the current observer was present when the cache had been made. For example, when bird A was present, cachers re-cached the items in the tray observed by A during the caching, but they did not touch the tray which contained food whose caching had not been observed by A. If the cacher can remember who was watching during their caching stage, all the different in re-caching can be explained. During the re-caching stage, when individual A was watching, the cacher only needs to recache the caches that are in the tray where caching had been observed by A. And the cacher should not recache the food in the other tray since A wouldn't know that there was food in it. The same strategy applied to the presence of B. But when the cacher was observed by a third bird which they had not encountered during the caching stage, the cachers should not touch any tray at all to expose the caching information to a stranger. What could the behaviour reading theory say to explain these results? I will try my best to think some arguments for them. Corvids can associate the specific observer with the place they cached the food. They may learn that only the place watched by a specific observer would lead to losing caches, and the other places with caches were just fine. So, when the corvids need to recache the food, they only recached items from the place related to the observer. The behaviour rule could be something like: only re-cache items from places associated with the particular conspecific present during caching. It is a more complex rule than some simple rules like re-cache food when a conspecific was present when the food was cached. But I think the behaviour reading theory has some grounds for attributing more complex rules of this type. They could learn that not every observing conspecific would lead to loss in caches. This would push them to find something more subtle like the identity of the observers as an indicator. At least, this experiment cannot rule out such behaviour rule. However, other cases can put a heavier explanatory burden on the behaviour reading theory.

Thomas Bugnyar (2011) designed another study, involving ravens, to show that corvids know who was watching them, and that they can adjust their behaviour based on others' knowledge. In his study, two observing ravens competed with a third raven, the subject, for food cached by a human

experimenter. During the caching stage, the experimenter placed two pieces of food into two caches between the birds. The subject raven had witnessed where both caches were, while each competitor was only allowed to see one cache (Fig 4.a). Later, the subject raven was given the advantage to recover just one cache in front of one of the competitors (Fig 4.b). Then the competitor got the chance to recover a cache. It is no surprise from both the mindreading and the behaviour reading theory that the subject raven tended to recover the cache paired with the observed competitor. This means that the subject would recover the cache known to the competitor it faced at the time. The mindreading theory explains this by saying that the subject attributes to each competitor knowledge of where either the first cache was or the second cache was, so the subject would first recover the cache it takes to be known by the current competitor. The behaviour reading theory would explain it by saying the subject raven associates the cache with the competitor who observed the caching stage. This is the same behaviour rule I have discussed in the previous case. But the genius of this experiment is that there was another setup.

Figure 4



Note. This picture was produced by Bugnyar in 2005, showing the sketch of the experimental setup during caching and testing. From "Knower-guesser differentiation in ravens: others' viewpionts matter," by T. Bugnyar, 2011, *Proceeding of the Royal Society*, 278(1705), p635 (https://doi.org/10.1098/rspb.2010.1514)

Figure 5

Note. This picture was produced by Bugnyar in 2005, showing focal subject's view of the competitor. From "Knower-guesser differentiation in ravens: others' viewpionts matter," by T. Bugnyar, 2011, *Proceeding of the Royal Society*, 278(1705), p637 (https://doi.org/10.1098/rspb.2010.1514)

In this setup, the competitor ravens were put in a specially designed room. In the room, the competitor raven perched at a higher location than the subject raven. An opaque curtain with a window on it can be pulled up or pulled down during the experiment. When the curtain was pulled up (Fig 5), the subject raven can still see the competitor, but the competitor cannot see where the cache was put because the caching place was too close to the door. During the caching stage, competitors were put into two conditions. In the informed condition, the curtain was pulled down, allowing competitors to see the caching stage as well as the subject. In the uninformed condition, the curtain was pulled up, so the competitors were not able to see where the food was placed. But in both conditions, the subject raven can always see the presence of the competitor. The result showed that the subject would only recover the cache in the informed condition but not in the uninformed condition. Let me reiterate the behaviour rule that successfully explained the results in previous setup. On that rule, the assumption was that when the observer was presented in the caching of the food, the food cached in the place associated with the observer was more likely be to lose if not re-cached. So, when an observer was present, subject ravens in the previous setup recovered the food from the place associated with that observer. This behaviour rule cannot work in this case. From the view of the subject raven, a competitor was associated with a caching place. So, whether the competitor can or cannot see the caching place should not matter. The subject raven should always recover the corresponding cache when a certain competitor was present. In this case, that means, in both conditions, the subject should always recover the food based on

which competitor was present. But this prediction made from the behaviour reading theory contradicts the results. The subject's recovery was based on the knowledge of the competitor instead of an associated relationship between the presence of the competitor and the cache. However, even if this result rules out this behaviour rule, the behaviour reading theory still has something more to offer. For example, it could argue that beyond the identity of the observers and the location associated with their presence, there is one more thing to consider, which is the place should be in their direct line of sight. So, only places that are in the direct line of sight of an observer would be associated with that observer. This rule has the potential to be learned. In their past experience, being in the direct line of sight of an observer could lead to an even bigger chance of caches being lost. So, the mere presence of an observer may not be a sufficient threat to lead them to re-cache their food. But I will now argue that even with this patch to the behaviour rule, the behaviour reading account still fails.

Bugnyar et al. (2016) designed an experiment to show that corvids can attribute knowledge to a conspecific not in their sight. I have already presented this study in chapter 1. Due to its significant role in this chapter, I would like to reintroduce it. This study will invalidate the additional behaviour rule proposed in the previous paragraph. No matter how hard the behaviour reading theory tries to patch their rules, there is always a loophole. In Bugnyar et al.'s experiment, there are two rooms divided by a wall, one for an observer and the other for a cacher. The wall has two separate windows which can be closed or opened. One is a big window which is big enough to expose the observer to the view of the cacher, and the other window is a peephole which enables the observer to see the caching behaviour while the observer is not visible to the cacher. There are three setups. (a)The observed condition, the big window is open. The caching raven caches the food under the attendance of the observing raven, and the caching bird can see the observer. (b)The non-observed condition, the big window is closed. The cacher caches the food without being observable by the observer. In both conditions, the raven in the observing room makes some noise. (c)The peephole condition, the big window is closed while the peephole is open. In this condition, when the cacher caches food there is no bird present in the observing room, but since the peephole is so small, the caching bird cannot see whether there is a bird at the peephole. And experimenters play the sound of a bird making noise in the observing room through a loudspeaker in that room (the recording played was of the bird in the observing room in the non-observed condition). If the

caching ravens have not experienced the view from the observing room through the peephole they may take no notice of it. However, once ravens have had the experience of observing other ravens' caching behaviours through the peephole, they will behave differently when they themselves are doing the caching. In the peephole condition, when ravens cached in storing room, they will cache more quickly than in the unobserved condition and they will be more likely return to re-cache the food. These caching behaviours are similar to the behaviour of ravens in the observed condition. Both were different from the unobserved condition, in which ravens took more time to cache and were less likely to revisit to improve their caches through re-caching.

The behaviour rule I have suggested before was that the corvids can associate who is observing them with a direct line of the sight. This rule gives corvids the ability to not only track who is watching them but also react to such observations with certain protection behaviours. In this case, this behaviour rule falls short of explaining the corvids' caching behaviour. In the peephole condition, the observer was not observed to have a direct line sight to the cacher (in fact no observer was present, there were only sounds suggesting the presence of an observer). Based on the behaviour rule, caching birds should not take any cache protection behaviour. There was only one obvious difference between the unseen condition and the peephole condition: the peephole was open in the peephole condition. Can corvids associate the opened peephole with the cache protection behaviour? One way to do so is to link the open peephole with observers having a direct sight. Since these ravens were never exposed to such peepholes prior to this experiment, it is very hard for them to draw such connections. The only experience they had with the peephole was in the familiarisation trials. Ravens were called by their name by the experimenters, and experimenters showed them the food in the next room. When the raven looked through the hole, the experimenter cached the food in the room. Only raven who were able to correctly recover the hidden food were included in this study. Since the familiarisation trials had nothing to do with cache pilferage, it is very hard to draw a direct connection between them. Certainly, the ravens had not learned through experience that when they cached food near to peepholes they were more likely to lose the cached food if the peepholes were open than if they were closed. It was also very hard to connect the peephole with an observer because ravens had never looked through the hole to locate an observer previously. From pure logic, there is nothing to stop an advocate of the

behaviour reading theory from positing such a connection. But in this case, I cannot see that they would have any justified reason to do so.

Based on the studies that I have presented, I think the pattern here is clear. The behaviour reading theory promises a simpler way to explain animal behaviours, but it ends up having to add more and more patches in their theory. The most obvious behavioural rules always end up being excluded by the control conditions. Moreover, there is no single behavioural rule that can explain corvids' caching behaviour once for all. In some cases, like that of Bugnyar et al. (2016), it is very hard to find any reasonable behavioural rule at all to explain the data.

Another way to try to save the behaviour reading theory is to come up with a more abstract rule. For example, a rule such as this: when you are observed by others during caching, remember who the observer is, and later, take action to protect the caches. The terms like "are observed" and "protect" are highly ambiguous in the eye of behavioural definition. Being observed is an abstract concept which can be related to many different behavioural cues such as being in the direct line of sight or being in the presence of an open peephole. These terms are also very closely linked with mentalistic interpretations. Being observable explicitly so, and protection implicitly, since protection behaviours depend on who observed and what they observed. It's worth mentioning the threat of the abstract rules smuggling in mentalistic elements. If corvids can learn concepts like this, the behaviour reading theory already gives up their most important argument which is that animals cannot learn unobservable concepts.

Since the behaviour reading theory has difficulty explaining the corvids' behaviour in what I am calling knowledge attribution, we should ask whether the minimal theory of mind account can deal with them better? Let me illustrate how the protective behaviour from corvids can be explained in the term of the minimal theory of mind account. When the cacher encounters an observer during their caching, he would assign the goal of pilfering his caches to the observer. To avoid having his caches pilfering, the cacher needs to take action to interfere with the observer's having a correct registration of the cache location. To achieve that he can choose to re-cache or eat the food, as long as his action leads to the observer having an incorrect registration of the cache location. This account closely parallels how a ToM account would explain the birds' behaviour, but such account

uses the minimal theory of mind's more minimal concepts of goals and registrations in place of the ToM accounts richer mentalistic concepts. To handle the cases that corvids can track the sound from both the pilferer's and the protector's' view, the minimal theory needs to add some principles. For example, to encounter an object, the attributer can not only take the visual field but also the auditory field in account. In the case of tracking an unseen observer, the minimal theory of mind account can provide a reasonable explanation. By registering the observer's last encountering, the attributer can track an unseen observer's perception in this way. At present, the cases of knowledge attribution in corvids can be explained by the minimal theory by some tuning. But since the primary target of the researchers who designed these studies with corvids was the behaviour reading theory, I would not be surprised to see the minimal theory fail in the case of corvids just like it does in the chimpanzee's studies, once researchers design studies with corvids parallel to those that have been done with chimpanzees.

Besides the mindreading theory and the minimal theory, there is another hypothesis that has been offered to try to explain why corvids behaved as they do in the cache protection proposed by Van der Vaart and colleagues (2012). They claimed that the protection behaviour such as re-caching was a side-effect of stress. The presence of a competitor would increase the stress level of the cacher. This increased stress level led to a general desire to cache more. It was suggested that the stress level of the cacher was influenced by factors such as how close the competitor was, the dominance relationship between them, and unsuccessful recovery attempts. Thom and Clayton (2013) designed a specific experiment to test this hypothesis. In their setup, ravens were put into a stress condition and a non-stress condition. In the stress condition, ravens were provided with peanuts to cache in a private cage, but all the caches were removed (that is, pilfered) by the experimenter out ravens' view. Then, the ravens returned to the cage. This time they were given new peanuts and a new tray alongside their old tray. In the non-stress condition, ravens were given an empty bowl and a tray in the private cage. Then they were given the same recovery condition as the stress condition. If the stress hypothesis worked, ravens in the stress condition should cache more because they had the experience of pilferage. But the results did not show such an effect. Raven in both conditions cached the same amount of peanuts. Moreover, in the previous sections, I have discussed the case that the caching birds can track who is observing them during the caching, and they only take protection behaviour when the same observer is present. The stress hypothesis

will have trouble to example such result because identity shouldn't matter for them as long as an observer is there.

As I have showed in this section, corvids can track what others can perceive. They can also track who is perceiving. Moreover, the subject corvids can combine the information about what others can hear and what they can see together to help them determine when they should re-cache items. For any single experiment, the behaviour reading account and the minimal theory account may come up a way to explain the evidence by adjusting their original theories. But when we consider the whole picture here, the mindreading theory is the most reasonable one.

3. Desire attribution

Besides the standard phenomena of attributions of knowledge and belief found in chimpanzees, studies of corvids provided employed another paradigm involving attributing mental states to other which is desire attribution. Desire attribution involves understanding what others want given their situation, and at least sometimes, the desires that are attributed to others may conflict with your own desires and preferences. This phenomenon of tracking others' desires had been reported in studies of corvids from food sharing contexts (Amodio et al., 2021; Ostojić et al., 2016; Ostojić et al., 2017; Ostojić et al., 2014; Ostojic et al., 2013) and food caching contexts (Ostojić et al., 2017)

First, let us look at the case in food sharing. Food sharing in corvids can be found in different situations. One of the most common forms is birds feeding their paired partners during courtship. Ostojić and colleagues (2013) studies how male Eurasian jays decided what to share. In their study, the males can choose two types of food to feed their partners: one was wax moth larvae (W), the other one is mealworm larvae (M). Before the sharing stage, the females ate either W, M, or a maintenance diet (MD), and the males always ate MD. Male jays can see what the females had been fed during this stage. Both male and female jays preferred a varied diet, and so having been fed only one type of food, they would then prefer a different type. Then the males had the chance to feed their partners from both W and M. The results showed that they shared a lower proportion of W when the females had been pre-fed only with W rather than pre-fed only with M. This means that male jays will share less W with their partner when she ate only W before than when she ate

only M. This result matched with jays' (both males and females) own preference for W or M in a previous satiety test (both preferred a change in diet). This means that the difference in sharing that the experimenters observed may be explained by males understanding the preferences of the females. If that is true, males can track their partner's preference in food sharing. Of course, instead of attributing desires or preferences to their partners, the corvids could use other tricks to achieve the same tracking ability. The researchers took two such alternative explanations into account. One was that the male may be getting a hint from the female during the food sharing. To address this possibility, the researchers introduced an unseen condition which was same the previous test except before the sharing stage, the males cannot see what females had been pre-fed. If the females were signalling to males about what they want, the males should still have access to these signals in the unseen condition, and the results should remain the same as in the previous condition. But the results did not support such a hypothesis. The significant sharing behaviour found in the seen condition disappeared. So, the tracking ability is not a simple reaction to the signal from the females. The other hypothesis the researcher took into consideration was called the "observational specific satiety" hypothesis. The idea was that male jays may themselves get vicariously satiated by watching what the female ate. In this case, when the male observed the female eat M, he also felt like he had been fed with M. So, their choice of sharing only reflected their own desires instead of other's desire. To address this hypothesis, the researchers added an extra experiment. The males watched the female eat the W, M, or MD just like in the seen condition described earlier. Then, the males were given the chance to eat from either W or M themselves. If the males behave as the observational specific satiety hypothesis suggests, they should eat in accordance with how they shared in the previous test. But the result showed they did not. The males didn't show a preference for the food the females were fed. It is also worth mentioning that Ostojić et al. (2014) conducted a further study of whether corvids can differentiate their own desires from the desires of others. By manipulating how both males and females were pre-fed, males and females can end up with conflicting desires in the food sharing stage. Many of the procedures were the same as the one I have introduced. The difference is that, during the testing phase, the males could either eat cache or feed the food to female. The results showed that the male's choices to eat are always in line with their only desire in different conditions, but their sharing choices are not. When the male's desired food type matched with the female's desired food type, the males were totally in line with their own choices to share. When the male's desire conflicted with that of the female, the males'

sharing pattern was different from their own choices to eat. This may because when the male held a contradicted desire between himself and his partner, his sharing behaviour is influenced by his female partner. This study solidified the claim that the male jays can distinguish their own desire from others to some degree.

In addition to the possible hypotheses we've considered so far in relation to food sharing, are there other plausible alternatives? First, let us consider the behaviour reading theory. The seen and unseen conditions in the study above do a very good job of ruling out behavioural cues from the female which may trigger particular aspects of the males' sharing behaviour. The females did not behave differently in the two conditions when their male partners fed them. The only visible difference between these two conditions came from the pre-feeding stage of the females (which was only observable by males in the seen condition). Could the males have learned the connection between what the female ate and what to feed purely on the basis of the behaviour they observed at this stage? It is possible. Providing a more desirable food during courting could win a potential partner. The males may learn to associate what the female ate with what foods are more successful in attracting and retaining potential mates in courting. Such behavioural rules could be something like: if the female ate A, offer her something else. But this rule has a flaw, what she ate may be what she really likes. Corvids had different preferences towards different types of food (and these preferences change in complex ways in response to different circumstances). As the experiment showed, the food shared by males changed based on what females ate changed in proportional terms not absolute numbers. For example, when the male sees that the female had been pre-fed W, the male will lower the proportion of W shared with his partner. But if we only consider the quantity, the male could still share more W more M. This result is contradicted by the proposed behavioural rule. Because the rule predicted the male would offer a lower proportion of W than other choices. The problem here for this behavioural rule is that the corvids behaved based on the relative proportion, but the behaviour rule is framed in terms of absolute quantities. To match with such proportion-based effects, the behavioural rule needs to propose that corvids can perceive the proportions of certain items. The rule would be like: offer more of M, or less of M based on how much M the female ate before. I'm not sure whether corvids are sensitive to number in a way like this. Even if corvids actually can perceive proportional information, the behaviour reading theory needs to provide evidence for that; it can't just be assumed. On the other side, the mindreading

approach can avoid such problems because the male can have a basic understanding of the female's desire. He could tell what she desired at the beginning and adjust how to feed her later based on her current desire.

The second area of research regarding desire attribution in corvids is centred around their cache protection. Ostojić and colleagues (2017) conducted another study about desire attribution focused on this. In this study, they first demonstrated that scrub-jays would adjust their pilferage based on what they ate before. When birds had been pre-fed with food A, they would prefer to pilfer food B relative to the baseline of their original preference for these two foods. Then, experimenters started the caching study. The procedure of this experiment was like the previous food sharing one. The difference was that the caching birds needed to cache food A or B after they have witnessed what the pilfers had been pre-fed, A or B. In the seen condition, the cacher can see what the pilferer ate, but not in the unseen condition. The results showed that cachers would cache more of the type of food that the pilferers had eaten compared to the baseline. This means the storer will cache more food A over food B relative to the baseline when they observed that the pilferer had eaten food A. This result is consistent with the ToM account, which would say that the birds are attributing desires. The behaviour reading theory suffers from the same difficulty as I have noted before. Namely how could the behaviour reading theory propose a rule to cover corvids' sensitivity to the pre-fed food they have observed.

The minimal theory of mind account shows its limitations in this context as well. Since this theory is mainly focused on how to track perception and knowledge, it is unsuited to explain how other mental states such as desires are tracked. Because the theory only allows the agent to track location information about items, it would need to be extended even to be able to track the perception and knowledge states involved in these studies. In these studies, the male feeders only noticed what type of food their partners had been given. To track such information, the minimal theory would need some additional principles like: "when a potential partner has only been able to eat one type of food, then avoid feeding them that same type of food". But such a principle is essentially just a behavioural rule of the sort I have given earlier, so the minimal theory of mind account, even if extended in this way, would still suffer from the same problems as the behaviour reading theory.

Finally, what should we say about belief attribution vs. knowledge attribution in corvids? The picture of mindreading in corvids is less conclusive compared to the case of chimpanzees. The main reason for this however is the fact that there are no false belief tasks that have been run using corvids. This makes it impossible to decide between the belief attribution theory (BAT) and the knowledge attribution theory (KAT). Based on current evidence in corvids, KAT does as well as BAT. But these two theories both agree that the corvids can attribute certain mental states like perception and knowledge to others. BAT and KAT should therefore be considered to be different versions of the mindreading theory. So, this means that regardless of what happens when KAT and BAT are pitted against one another, a mindreading theory is still favoured in the case of corvids comparing to non-mindreading theory like the behaviour reading and minimal theory.

Summary

In this chapter, I have argued that corvids, just like chimpanzees, are capable of attributing mental states to others just as the mindreading theory suggests. I have followed the same methodology as I employed in the previous chapter on chimpanzees to check how each theory does in trying to explain the current experimental evidence. And as in the case of chimpanzees, the most consistent and persuasive theory among them is the mindreading theory. But since the studies of corvids are still lacking data on how corvids perform in the false belief tests, I cannot make the call that the belief attribution theory (the standard mindreading theory) is better than the knowledge attribution theory in the case of corvids. But based on the reasons I have presented in chapter 3 about why the knowledge attribution theory is a less desirable theory between the two, I am still inclined to take the mindreading theory to be the best theory to explain corvids' behaviour.

Conclusion

In this thesis, I argue that the mindreading theory is the most reasonable explanation of the behaviour of animals like chimpanzees and corvids, both as found in the wild and in the experiments. As always the case in scientific studies, any single piece of evidence found in the experiments is shorted on proofing any theory. Instead of listing positive evidence for supporting the mindreading theory, my strategy in this thesis is to compare the mindreading theory with three other promising and popular alternative theories. These three theories are not chosen by random. Instead, each of them can be considered as a representative of a type of theories.

The first alternative theory is the behaviour reading theory (BRT) which is the traditional approach to study animals and even human beings. The idea is that to explain animals' performance, we don't need anything more than behaviour cues. Since its huge success on explaining many other animal behaviours, it is often considered as a default model to understand all of their behaviours. But as I argued in chapter 1, the theory itself is problematic. It lacks a concrete explanation about how the theory works. If we think about BRT in terms of behaviour rules, any specific rules failed to do the job. A more flexible approach of thinking behaviour rules faces similar criticisms if not even more problems than the mindreading approach.

The second alternative theory, the minimal theory, aims to replace mental states like belief with something less cognitively demanding. The minimal theory hopes to build a theory with minimal mentalistic concepts and principles to deliver a similar function as that which the mindreading theory promises. The problem with this approach is that it is much harder than they think to create an alternative way to mimic folk psychology functions. To fully appreciate what animal can do in experiments, they have to add more principles than they suggested. But such ad hoc way to add more functions by itself lacking predictive power for any further findings. Moreover, even by adding some principles as they suggested, the theory still fall short in some cases.

The third theory is the knowledge attribution theory(KAT), which shares a lot with the mindreading theory. I would say it is a version of the standard mindreading theory, or more specific the belief attribution theory (BAT). I'm not really arguing against it, rather I argue that KAT is a

less desirable version of BAT. By placing knowledge as the default mental states instead of belief when individual explain other's behaviour, the theory claims it is better suited for the evidence. The only noticeable difference between them is whether individuals can attribute belief to others. KAT claims animals can only attribute knowledge but not belief to others. But if we take knowledge as the form of true beliefs, most evidence they provided to support KAT is equally supporting BAT. The only evidence can really distinguish KAT and BAT is the cases they provided to show apes and monkeys can attribute knowledge but not true beliefs in certain experimental studies. But such evidence, as I have argued in chapter 3, does not work in their favour as well. Besides, more and more promising results in false belief test came out recently in chimpanzee's studies. Such evidence undermines the original motivation of proposing a knowledge attribution theory.

Since each of these approaches has their own problems, I turn my effort to draw a whole picture how the mindreading theory is the best available theory to explain the behaviour of both chimpanzees and corvids in the following two chapters. In these two chapters, I focus on presenting the experimental evidence to support the idea that both chimpanzees and corvids can attribute mental states to others. Such mental states include goals and intentions, perception and knowledge. Based on the recent development on false belief tests in chimpanzees, it provided promising evidence that chimpanzees can understand what others believe even they are different from their own beliefs. For corvids, no false belief test has done yet due to their different morphology. But corvids show some other interesting abilities such as the ability to understand others' desires. Certainly, there are still questions remaining. For chimpanzees, why did they fail the previous false belief test but not the new ones? For corvids, is there a way to design their version of false belief test? All of these questions need further research.

Bibliography

Amodio, P., Farrar, B. G., Krupenye, C., & Ostojic…, L. (2021). Little evidence that Eurasian jays protect their caches by responding to cues about a conspecific's desire and visual perspective. *Elife*, *10:e69647*. https://doi.org/10.7554/eLife.69647

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states. *Psychological review*, *116*(4), 953. https://doi.org/10.1037/a0016923

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind". *Cognition*, *21*(1), 37-46. https://doi.org/10.1016/0010-0277(85)90022-8

Barone, P., Corradi, G., & Gomila, A. (2019). Infants' performance in spontaneous-response false belief tasks: A review and meta-analysis. *Infant Behavior and Development*, *57*. https://doi.org/10.1016/j.infbeh.2019.101350

Behne, T., Carpenter, M., Call, J., & Tomasello, M. (2005). Unwilling versus unable: infants' understanding of intentional action. *Developmental psychology*, *41*(2), 328-337. https://psycnet.apa.org/doi/10.1037/0012-1649.41.2.328

Bird, C. D., & Emery, N. J. (2009a). Rooks use stones to raise the water level to reach a floating worm. *Current Biology*, *19*(16), 1410-1414. https://doi.org/10.1016/j.cub.2009.07.033

Bird, C. D., & Emery, N. J. (2009b). Insightful problem solving and creative tool modification by captive nontool-using rooks. *Proceedings of the National Academy of Sciences*, *106*(25), 10370-10375. https://doi.org/10.1073/pnas.0901008106

Boesch, C., & Boesch, H. (1990). Tool use and tool making in wild chimpanzees. *Folia primatologica*, *54*(1-2), 86-99. https://doi.org/10.1159/000156428

Bossema, I. (1979). Jays and oaks: an eco-ethological study of a symbiosis. *Behaviour*, *70*(1-2), 1-116. https://doi.org/10.1163/156853979X00016

Bräuer, J., Call, J., & Tomasello, M. (2005). All great ape species follow gaze to distant locations and around barriers. *Journal of Comparative Psychology*, *119*(2), 145-154. https://psycnet.apa.org/doi/10.1037/0735-7036.119.2.145

Bräuer, J., Call, J., & Tomasello, M. (2007). Chimpanzees really know what others can see in a competitive situation. *Animal cognition*, *10*(4), 439-448. https://doi.org/10.1007/s10071-007-0088-1

Brown, C. (2012). Tool use in fishes. *Fish and Fisheries*, *13*(1), 105-115. https://doi.org/10.1111/j.1467-2979.2011.00451.x

Bugnyar, T., & Heinrich, B. (2005). Ravens, Corvus corax, differentiate between knowledgeable and ignorant competitors. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1573), 1641-1646. https://doi.org/10.1098/rspb.2005.3144

Bugnyar, T. (2011). Knower-guesser differentiation in ravens: others' viewpoints matter. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1705), 634-640. https://doi.org/10.1098/rspb.2010.1514

Bugnyar, T., & Heinrich, B. (2006). Pilfering ravens, Corvus corax, adjust their behaviour to social context and identity of competitors. *Animal Cognition*, *9*(4), 369-376. https://doi.org/10.1007/s10071-006-0035-6

Bugnyar, T., Reber, S. A., & Buckner, C. (2016). Ravens attribute visual access to unseen competitors. *Nature Communications*, *7*(10506). https://doi.org/10.1038/ncomms10506

Burt, D. W., Bruley, C., Dunn, I. C., Jones, C. T., Ramage, A., Law, A. S., Morrice, D. R., Paton, I. R., Smith, J., Windsor, D., Sazanov, A., Fries, R., & Waddington, D. (1999). The

dynamics of chromosome evolution in birds and mammals. *Nature*, *402*(6760), 411-413. https://doi.org/10.1038/46555

Buttelmann, D., Buttelmann, F., Carpenter, M., & Call…, J. (2017). Great apes distinguish true from false beliefs in an interactive helping task. *PLoS ONE*, *12*(4), e0173793. https://doi.org/10.1371/journal.pone.0173793

Buttelmann, D., Carpenter, M., Call, J., & Tomasello, M. (2007). Enculturated chimpanzees imitate rationally. *Developmental science*, *10*(4), F31-F38. https://doi.org/10.1111/j.1467-7687.2007.00630.x

Buttelmann, D., Schütte, S., Carpenter, M., Call, J., & Tomasello, M. (2012). Great apes infer others' goals based on context. *Animal Cognition*, *15*(6), 1037-1053. https://doi.org/10.1007/s10071-012-0528-4

Butterfill, S. A., & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind & Language*, *28*(5), 606-637. https://doi.org/10.1111/mila.12036

Call, J., Hare, B., Carpenter, M., & Tomasello, M. (2004). 'Unwilling' versus 'unable': chimpanzees' understanding of human intentional action. *Developmental Science*, *7*(4), 488-498. https://doi.org/10.1111/j.1467-7687.2004.00368.x

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, *12*(5), 187-192. https://doi.org/10.1016/j.tics.2008.02.010

Call, J., & Tomasello, M. (1998). Distinguishing intentional from accidental actions in orangutans (Pongo pygmaeus), chimpanzees (Pan troglodytes) and human children (Homo sapiens). *Journal of Comparative Psychology*, *112*(2), 192-206. https://psycnet.apa.org/doi/10.1037/0735-7036.112.2.192

Call, J., & Tomasello, M. (1999). A nonverbal false belief task: The performance of children and great apes. *Child development*, *70*(2), 381-395. https://doi.org/10.1111/1467-8624.00028

Call, J., Hare, B. A., & Tomasello, M. (1998). Chimpanzee gaze following in an object-choice task. *Animal cognition*, *1*(2), 89-99. https://doi.org/10.1007/s100710050013

Canteloup, C., & Meunier, H. (2017). 'Unwilling' versus 'unable': Tonkean macaques' understanding of human goal-directed actions. *PeerJ*, *5*, e3227. https://doi.org/10.7717/peerj.3227

Carpenter, M., Akhtar, N., & Tomasello, M. (1998). Fourteen-through 18-month-old infants differentially imitate intentional and accidental actions. *Infant behavior and development*, *21*(2), 315-330. https://doi.org/10.1016/S0163-6383(98)90009-1

Chomsky, N. (1959). On certain formal properties of grammars. *Information and control*, *2*, 137-167. https://doi.org/10.1016/S0019-9958(59)90362-6

Chomsky, N. (2013). *4. A Review of BF Skinner's Verbal Behavior*. Harvard University Press. https://doi.org/10.4159/harvard.9780674594623.c6/html

Clary Dawson, K. D. M. (2016). Clark's Nutcrackers (Nucifraga columbiana) Flexibly Adapt Caching Behavior to a Cooperative Context. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01643

Clayton, N. (2007). Animal cognition: crows spontaneously solve a metatool task. *Current Biology*, *17*(20), R894-5. https://doi.org/10.1016/j.cub.2007.08.028

Crockford, C., Wittig, R. M., Mundry, R., & Zuberbühler, K. (2012). Wild chimpanzees inform ignorant group members of danger. *Current Biology*, *22*(2), 142-146. https://doi.org/10.1016/j.cub.2011.11.053

Crockford, C., Wittig, R. M., & Zuberbühler, K. (2017). Vocalizing in chimpanzees is influenced by social-cognitive processes. *Science Advances*, *3*(11), e1701742. https://doi.org/10.1126/sciadv.1701742

Dally, J. M., Emery, N. J., & Clayton, N. S. (2006). Food-caching western scrub-jays keep track of who was watching when. *Science*, *312*(5780), 1662-1665. https://doi.org/10.1126/science.1126539

Dally, J. M., Emery, N. J., & Clayton, N. S. (2004). Cache protection strategies by western scrub–jays (Aphelocoma californica): hiding food in the shade. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *271*(suppl_6). https://doi.org/10.1098/rsbl.2004.0190

Dally, J. M., Emery, N. J., & Clayton, N. S. (2005). Cache protection strategies by western scrub-jays, Aphelocoma californica: implications for social cognition. *Animal Behaviour*, *70*(6), 1251-1263. https://doi.org/10.1016/j.anbehav.2005.02.009

Dally, J. M., Emery, N. J., & Clayton, N. S. (2010). Avian Theory of Mind and counter espionage by food-caching western scrub-jays (Aphelocoma californica). *European Journal of Developmental Psychology*, *7*(1), 17-37. https://doi.org/10.1080/17405620802571711

Dennett, D. C. (1978). Beliefs about beliefs [P&W, SR&B]. *Behavioral and Brain sciences*, *1*(4), 568-570. https://doi.org/10.1017/S0140525X00076664

Dinets, V., Brueggen, J. C., & Brueggen, J. D. (2015). Crocodilians use tools for hunting. *Ethology Ecology & Evolution*, *27*(1), 74-78. https://doi.org/10.1080/03949370.2013.858276

Drayton, L. A., & Santos, L. R. (2018). What do monkeys know about others' knowledge. *Cognition*, *170*, 201-2018. https://doi.org/10.1016/j.cognition.2017.10.004

Egyed, K., Király, I., & Gergely, G. (2013). Communicating shared knowledge in infancy. *Psychological Science*, *24*(7), 1348-1353. https://doi.org/10.1177%2F0956797612471952

Emery, N. J., & Clayton, N. S. (2001). Effects of experience and social context on prospective caching strategies by scrub jays. *Nature*, *414*(6862), 443-446. https://doi.org/10.1038/35106560

Emery, N. J., & Clayton, N. S. (2004). The mentality of crows: convergent evolution of intelligence in corvids and apes. *Science*, *306*(5703), 1903-1907. https://doi.org/10.1126/science.1098410

Emery, N. J., Seed, A. M., von Bayern, A. M. P., & Clayton, N. S. (2007). Cognitive adaptations of social bonding in birds. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 489-505. https://doi.org/10.1098/rstb.2006.1991

Emery, N. J., Dally, J. M., & Clayton, N. S. (2004). Western scrub-jays (Aphelocoma californica) use cognitive strategies to protect their caches from thieving conspecifics. *Animal Cognition*, *7*(1), 37-43. https://doi.org/10.1007/s10071-003-0178-7

Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the Level 1–Level 2 distinction. *Developmental Psychology*, *17*(1), 99-103. https://psycnet.apa.org/doi/10.1037/0012-1649.17.1.99

Flavell, J. H., Flavell, E. F., Green, F. L., & Wilcox, S. A. (1980). Young children's knowledge about visual perception: Effect of observer's distance from target on perceptual clarity of target. *Developmental Psychology*, *16*(1), 10. https://psycnet.apa.org/record/1980-05159-001

Flavell, J. H., Shipstead, S. G., & Croft, K. (1978). Young children's knowledge about visual perception: Hiding objects from others. *Child Development*, *49*(4), 1208-1211. https://doi.org/10.2307/1128761

Flombaum, J. I., Junge, J. A., & Hauser, M. D. (2005). Rhesus monkeys (Macaca mulatta) spontaneously compute addition operations over large numbers. *Cognition*, *97*(3), 315-325. https://doi.org/10.1016/j.cognition.2004.09.004

Frege, G. (1948). Sense and reference. *The philosophical review*, *57*(3), 209-230. https://www.jstor.org/stable/2181485

Gallagher, S., & Povinelli, D. J. (2012). Enactive and behavioral abstraction accounts of social understanding in chimpanzees, infants, and adults. *Review of Philosophy and Psychology*, *3*(1), 145-169. https://doi.org/10.1007/s13164-012-0093-4

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, *2*(12), 493-501. https://doi.org/10.1016/S1364-6613(98)01262-5

Gergely, G., Bekkering, H., & Király, I. (2002). Rational imitation in preverbal infants. *Nature*, *415*(6873), 755-755. https://doi.org/10.1038/415755a

Goldman, A. I. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford University Press on Demand. https://doi.org/10.1093/0195138929.001.0001

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain sciences*, *16*(1), 1-14. doi:10.1017/S0140525X00028636

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. The MIT Press.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language*, *7*(12), 145-171. https://psycnet.apa.org/doi/10.1111/j.1468-0017.1992.tb00202.x

Graham, G. (2000). Behaviorism. *The Stanford Encyclopedia of Philosophy*, Spring 2019 Edition). https://plato.stanford.edu/archives/spr2019/entries/behaviorism/

Graham, K. E., Wilke, C., Lahiff, N. J., & Slocombe, K. E. (2020). Scratching beneath the surface: intentionality in great ape signal production. *Philosophical Transactions B*, *375*(1789), 20180403. https://doi.org/10.1098/rstb.2018.0403

Gruber, R., Schiestl, M., Boeckle, M., Frohnwieser, A., Miller, R., Gray, R. D., Clayton, N. S., & Taylor, A. H. (2019). New Caledonian Crows Use Mental Representations to Solve Metatool Problems. *Current Biology*, *29*(4), 686-692.e3. https://doi.org/10.1016/j.cub.2019.01.008

Hare, Call, Agnetta, & Tomasello. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour*, *59*(4), 771-785. https://doi.org/10.1006/anbe.1999.1377

Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know. *Animal Behaviour*, *61*(1), 139-151. https://doi.org/10.1006/anbe.2000.1518

Hare, B., Call, J., & Tomasello, M. (2006). Chimpanzees deceive a human competitor by hiding. *Cognition*, *101*(3), 495-514. https://doi.org/10.1016/j.cognition.2005.01.011

Hauser, M. D., & Wrangham, R. W. (1987). Manipulation of food calls in captive chimpanzees: A preliminary report. *Folia primatologica*, *48*(3-4), 207-210. https://psycnet.apa.org/doi/10.1159/000156298

Hayashi, T., Akikawa, R., Kawasaki, K., Egawa, J., Minamimoto, T., Kobayashi, K., Kato, S., Hori, Y., Nagai, Y., & Iijima, A. (2020). Macaques exhibit implicit gaze bias anticipating

others' false-belief-driven actions via medial prefrontal cortex. *Cell reports*, *30*(13), 4433-4444. e5. https://doi.org/10.1016/j.celrep.2020.03.013
Heyes, C. (2017). Apes submentalise. *Trends in Cognitive Sciences*, *21*(1), 1-2. https://doi.org/10.1016/j.tics.2016.11.006
Heyes, C. (1998). Theory of mind in nonhuman primates. *Behavioral and Brain Sciences*, *21*(1), 101-114. https://doi.org/10.1017/s0140525x98000703
Heyes, C. (2014). Submentalizing: I am not really reading your mind. *Perspectives on Psychological Science*, *9*(2), 131-143. https://doi.org/10.1177/1745691613518076
Heyes, C. (2015). Animal mindreading: what's the problem. *Psychonomic bulletin & review*, *22*(2), 313-327. https://link.springer.com/article/10.3758/s13423-014-0704-4
Horschler, D. J., MacLean, E. L., & Santos, L. R. (2020). Do non-human primates really represent others' beliefs. *Trends in Cognitive Sciences*, *24*(8), 594-605. https://doi.org/10.1016/j.tics.2020.05.009
Horschler, D. J., Santos, L. R., & MacLean, E. L. (2019). Do non-human primates really represent others' ignorance? A test of the awareness relations hypothesis. *Cognition*, *190*, 72-80. https://doi.org/10.1016/j.cognition.2019.04.012
Horschler, D. J., Santos, L. R., & MacLean, E. L. (2021). How do non-human primates represent others' awareness of where objects are hidden. *Cognition*, *212*, 104658. https://doi.org/10.1016/j.cognition.2021.104658
Hunt, G. R. (1996). Manufacture and use of hook-tools by New Caledonian crows. *Nature*, *379*(6562), 249-251. https://doi.org/10.1038/379249a0
Hunt, G. R. (2000). Human–like, population–level specialization in the manufacture of pandanus tools by New Caledonian crows Corvus moneduloides. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *267*(1441), 403-413. https://doi.org/10.1098/rspb.2000.1015
Hunt, G. R., & Gray, R. D. (2004). The crafting of hook tools by wild New Caledonian crows. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *271*(3), S88-90. https://doi.org/10.1098/rsbl.2003.0085
Hurley, S. E., & Nudds, M. E. (2006). *Rational animals*. Oxford University Press. https://psycnet.apa.org/record/2006-08631-000
Jacobs, I. F., von Bayern, A., Martin-Ordas, G., Rat-Fischer, L., & Osvath, M. (2015). Corvids create novel causal interventions after all. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1806), 20142504. https://doi.org/10.1098/rspb.2014.2504
Jerison, H. J. (1973). Evolution of the Brain in Birds. In *Evolution of the Brain and Intelligence* (pp. 177-199). Academic Press. https://doi.org/10.1016/b978-0-12-385250-2.50018-3
Kabadayi, C., & Osvath, M. (2017). Ravens parallel great apes in flexible planning for tool-use and bartering. *Science*, *357*(6347), 202-204. https://doi.org/10.1126/science.aam8138
Kaminski, J., Call, J., & Tomasello, M. (2008). Chimpanzees know what others know, but not what they believe. *Cognition*, *109*(2), 224-234. https://doi.org/10.1016/j.cognition.2008.08.010
Kano, F., Krupenye, C., Hirata, S., Tomonaga, M., & Call, J. (2019). Great apes use self-experience to anticipate an agent's action in a false-belief test. *Proceedings of the National Academy of Sciences*, *116*(42), 20904-20909. https://doi.org/10.1073/pnas.1910095116
Karg, K., Schmelz, M., Call, J., & Tomasello, M. (2016). Differing views: Can chimpanzees do Level 2 perspective-taking. *Animal Cognition*, *19*(3), 555-564. https://doi.org/10.1007/s10071-016-0956-7

Krachun, C., Call, J., & Tomasello, M. (2010). A new change-of-contents false belief test: Children and chimpanzees compared. *International Journal of Comparative Psychology*, *23*(2), 145-165. https://escholarship.org/uc/item/68c0p8dk

Krachun, C., Carpenter, M., Call, J., & Tomasello, M. (2009). A competitive nonverbal false belief task for children and apes. *Developmental science*, *12*(4), 521-535. https://doi.org/10.1111/j.1467-7687.2008.00793.x

Krupenye, C., & Call, J. (2019). Theory of mind in animals: Current and future directions. *WIREs Cognitive Science*, *10*(6), e1503. https://doi.org/10.1002/wcs.1503

Krupenye, C., Kano, F., Hirata, S., & Call…, J. (2017). A test of the submentalizing hypothesis: Apes' performance in a false belief task inanimate control. *Communicative & Integrative Biology*, *10*(4), e1343771. https://doi.org/10.1080/19420889.2017.1343771

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, *354*(6308), 110-114. https://doi.org/10.1126/science.aaf8110

Legg, E. W., & Clayton, N. S. (2014). Eurasian jays (Garrulus glandarius) conceal caches from onlookers. *Animal Cognition*, *17*(5), 1223-1226. https://doi.org/10.1007/s10071-014-0743-2

Luo, Y., & Beck, W. (2010). Do you see what I see? Infants' reasoning about others' incomplete perceptions. *Developmental Science*, *13*(1), 134-142. https://doi.org/10.1111/j.1467-7687.2009.00863.x

Lurz, R. (2009). If chimpanzees are mindreaders, could behavioral science tell? Toward a solution of the logical problem. *Philosophical Psychology*, *22*(3), 305-328. https://doi.org/10.1080/09515080902970673

Lurz, R. W., & Krachun, C. (2011). How Could We Know Whether Nonhuman Primates Understand Others' Internal Goals and Intentions? Solving Povinelli's Problem. *Review of Philosophy and Psychology*, *2*(3), 449-481. https://doi.org/10.1007/s13164-011-0068-x

Marticorena, D. C. W., Ruiz, A. M., Mukerji, C., Goddu, A., & Santos, L. R. (2011). Monkeys represent others' knowledge but not their beliefs. *Developmental science*, *14*(6), 1406-1416. https://doi.org/10.1111/j.1467-7687.2011.01085.x

Martin, A., & Santos, L. R. (2016). What Cognitive Representations Support Primate Theory of Mind. *Trends in Cognitive Science*, *20*(5), 375-382. https://doi.org/10.1016/j.tics.2016.03.005

Melis, A. P., Call, J., & Tomasello, M. (2006). Chimpanzees (Pan troglodytes) conceal visual and auditory information from others. *Journal of Comparative Psychology*, *120*(2), 154-162. https://psycnet.apa.org/doi/10.1037/0735-7036.120.2.154

Melis, A. P., Warneken, F., Jensen, K., Schneider, A.-C., Call, J., & Tomasello, M. (2011). Chimpanzees help conspecifics obtain food and non-food items. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1710), 1405-1413. https://doi.org/10.1098/rspb.2010.1735

Miyamoto Gómez, O. S. (2021). Four Epistemological Gaps in Alloanimal Episodic Memory Studies. *Biosemiotics*, *14*, 839-857. https://doi.org/10.1007/s12304-021-09437-9

Moll, H., & Meltzoff, A. N. (2011). How does it look? Level 2 perspective-taking at 36 months of age. *Child development*, *82*(2), 661-673. https://doi.org/10.1111/j.1467-8624.2010.01571.x

Nagel, J. (2017). Factive and nonfactive mental state attribution. *Mind & Language*, *32*(5), 525-544. https://doi.org/10.1111/mila.12157

Oakley, K. P. (1964). *Man the tool-maker*. University of Chicago Press.
Okamoto-Barth, S., Call, J., & Tomasello, M. (2007). Great Apes' Understanding of Other Individuals' Line of Sight. Psychological Science, 18(5), 462-468. https://doi.org/10.1111/j.1467-9280.2007.01922.xS
Olineck, K. M., & Poulin-Dubois, D. (2005). Infants' ability to distinguish between intentional and accidental actions and its relation to internal state language. *Infancy*, *8*(1), 91-100. https://doi.org/10.1207/s15327078in0801_6
Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs. *science*, *308*(5719), 255-258. https://doi.org/10.1126/science.1107621
Ostojić, L., Cheke, L. G., Shaw, R. C., Legg, E. W., & Clayton, N. S. (2016). Desire-state attribution: Benefits of a novel paradigm using the food-sharing behavior of Eurasian jays (Garrulus glandarius). *Commun Integr Biol*, *9*(2), e1134065. https://doi.org/10.1080/19420889.2015.1134065
Ostojić, L., Legg, E. W., Brecht, K. F., Lange, F., Deininger, C., Mendl, M., & Clayton, N. S. (2017). Current desires of conspecific observers affect cache-protection strategies in California scrub-jays and Eurasian jays. *Current Biology*, *27*(2), R51-R53. https://doi.org/10.1016/j.cub.2016.11.020
Ostojić, L., Legg, E. W., Shaw, R. C., Cheke, L. G., Mendl, M., & Clayton, N. S. (2014). Can male Eurasian jays disengage from their own current desire to feed the female what she wants. *Biology Letters*, *10*(3), 20140042. https://doi.org/10.1098/rsbl.2014.0042
Ostojic, L., Shaw, R. C., Cheke, L. G., & Clayton, N. S. (2013). Evidence suggesting that desire-state attribution may govern food sharing in Eurasian jays. *Proceedings of the National Academy of Sciences*, *110*(10), 4123-4128. https://doi.org/10.1073/pnas.1209926110
Ozonoff, S., & McEvoy, R. E. (1994). A longitudinal study of executive function and theory of mind development in autism. *Development and psychopathology*, *6*(3), 415-431. doi:10.1017/S0954579400006027
Pause, B. M., Zlomuzica, A., Kinugawa, K., Mariani, J., Pietrowsky, R., & Dere, E. (2013). Perspectives on Episodic-Like and Episodic Memory. *Frontiers in Behavioral Neuroscience*, *7*. https://doi.org/10.3389/fnbeh.2013.00033
Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Science*, *31*(2), 109-30; discussion 130. https://doi.org/10.1017/S0140525X08003543
Penn, D. C., & Povinelli, D. J. (2007a). Causal cognition in human and nonhuman animals: a comparative, critical review. *Annual Review of Psychology*, *58*, 97-118. https://doi.org/10.1146/annurev.psych.58.110405.085555
Penn, D. C., & Povinelli, D. J. (2007b). 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. https://doi.org/10.1098/rstb.2006.2023
Phillips, J., Buckwalter, W., Cushman, F., Friedman, O., Martin, A., Turri, J., Santos, L., & Knobe, J. (2021). Knowledge before belief. *Behavioral and Brain Sciences*, *44*, E140. doi:10.1017/S0140525X20000618
Phillips, J., & Norby, A. (2019). Factive theory of mind. *Mind & Language*, *36*(3), 3-26. https://doi.org/10.1111/mila.12267
Phillips, W., Barnes, J. L., Mahajan, N., Yamaguchi, M., & Santos, L. R. (2009). 'Unwilling' versus 'unable': capuchin monkeys' (Cebus apella) understanding of human intentional

action. *Developmental Science*, *12*(6), 938-945. https://doi.org/10.1111/j.1467-7687.2009.00840.x

Pillow, B. H., & Flavell, J. H. (1986). Young children's knowledge about visual perception: Projective size and shape. *Child Development*, *57*(1), 125-135. https://doi.org/10.2307/1130644

Povinelli, D. J., Eddy, T. J., Hobson, R. P., & Tomasello, M. (1996). What young chimpanzees know about seeing. *Monographs of the society for research in child development*, *61*(3), i-189. https://doi.org/10.2307/1166159

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), 203-210. https://psycnet.apa.org/doi/10.1037/0735-7036.104.3.203

Povinelli, D. J., & Vonk, J. (2004). We don't need a microscope to explore the chimpanzee's mind. *Mind & Language*, *19*(1), 1-28. https://doi.org/10.1111/j.1468-0017.2004.00244.x

Povinelli, D. J., & Vonk, J. (2003). Chimpanzee minds: suspiciously human. *Trends in Cognitive Science*, *7*(4), 157-160. https://doi.org/10.1016/s1364-6613(03)00053-6

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind. *Behavioral and Brain Sciences*, *1*(4), 515-526. https://doi.org/10.1017/s0140525x00076512

Raby, C. R., Alexis, D. M., Dickinson, A., & Clayton, N. S. (2007). Planning for the future by western scrub-jays. *Nature*, *445*(7130), 919-921. https://doi.org/10.1038/nature05575

Russell, J., Mauthner, N., Sharpe, S., & Tidswell, T. (1991). The 'windows task'as a measure of strategic deception in preschoolers and autistic subjects. *British journal of developmental psychology*, *9*(2), 331-349. https://doi.org/10.1111/j.2044-835X.1991.tb00881.x

Rutz, C., Klump, B. C., Komarczyk, L., Leighton, R., Kramer, J., Wischnewski, S., Sugasawa, S., Morrissey, M. B., James, R., St Clair, J. J. H., Switzer, R. A., & Masuda, B. M. (2016). Discovery of species-wide tool use in the Hawaiian crow. *Nature*, *537*(7620), 403-407. https://doi.org/10.1038/nature19103

Santos, L. R., Nissen, A. G., & Ferrugia, J. A. (2006). Rhesus monkeys, Macaca mulatta, know what others can and cannot hear. *Animal Behaviour*, *71*(5), 1175-1181. https://doi.org/10.1016/j.anbehav.2005.10.007

Savage-Rumbaugh, E. S., Rumbaugh, D. M., & Boysen, S. (1978). Sarah's problems of comprehension. *Behavioral and Brain Sciences*, *1*(4), 555-557. https://doi.org/10.1017/S0140525X0007655X

Schel, A. M., Townsend, S. W., Machanda, Z., Zuberbühler, K., & Slocombe, K. E. (2013). Chimpanzee alarm call production meets key criteria for intentionality. *PLoS ONE*, *8*(10), e76674. https://doi.org/10.1371/journal.pone.0076674

Schmelz, M., Call, J., & Tomasello, M. (2011). Chimpanzees know that others make inferences. *Proceedings of the National Academy of Sciences*, *108*(7), 3077-3079. https://doi.org/10.1073/pnas.1000469108

Scott, R. M., Baillargeon, R., Song, H. J., & Leslie, A. M. (2010). Attributing false beliefs about non-obvious properties at 18 months. *Cogntive Psychology*, *61*(4), 366-395. https://doi.org/10.1016/j.cogpsych.2010.09.001

Scott, R. M., Richman, J. C., & Baillargeon, R. (2015). Infants understand deceptive intentions to implant false beliefs about identity: New evidence for early mentalistic reasoning. *Cognitive Psychology*, *82*, 32-56. https://doi.org/10.1016/j.cogpsych.2015.08.003

Seed, A. M., Hanus, D., & Call, J. (2011). Causal Knowledge In Corvids, Primates and Children: More Than Meets The Eye? In T. McCormack, C. Hoerl, & S. Butterfill (Eds.). *Tool use and causal cognition*, 89-110. http://hdl.handle.net/11858/00-001M-0000-000F-F4A1-3

Setoh, P., Scott, R. M., & Baillargeon, R. (2016). Two-and-a-half-year-olds succeed at a traditional false-belief task with reduced processing demands. *Proceedings of the National Academy of Sciences*, *113*(47), 13360-13365. https://doi.org/10.1073/pnas.1609203113

Shaw, R. C., & Clayton, N. S. (2012). Eurasian jays, Garrulus glandarius, flexibly switch caching and pilfering tactics in response to social context. *Animal Behaviour*, *84*(5), 1191-1200. https://doi.org/10.1016/j.anbehav.2012.08.023

Shaw, R. C., & Clayton, N. S. (2013). Careful cachers and prying pilferers: Eurasian jays (Garrulus glandarius) limit auditory information available to competitors. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1752), 20122238. https://doi.org/10.1098/rspb.2012.2238

Shaw, R. C., & Clayton, N. S. (2014). Pilfering Eurasian jays use visual and acoustic information to locate caches. *Animal Cognition*, *17*(6), 1281-1288. https://doi.org/10.1007/s10071-014-0763-y

Skinner, B. F. (1963). Behaviorism at fifty. *Science*, *140*(3570), 951-958. https://www.jstor.org/stable/1711326

Skinner, B. F. (1985). Cognitive science and behaviourism. *British Journal of psychology*, *76*(3), 291-301. https://doi.org/10.1111/j.2044-8295.1985.tb01953.x

Smirnova, A., Zorina, Z., Obozova, T., & Wasserman, E. (2015). Crows spontaneously exhibit analogical reasoning. *Current Biology*, *25*(2), 256-260. https://doi.org/10.1016/j.cub.2014.11.063

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological science*, *18*(7), 587-592. https://doi.org/10.1111%2Fj.1467-9280.2007.01944.x

Stulp, G., Emery, N. J., Verhulst, S., & Clayton, N. S. (2009). Western scrub-jays conceal auditory information when competitors can hear but cannot see. *Biology Letters*, *5*, 583-585. https://doi.org/10.1098/rsbl.2009.0330

Suddendorf, T., & Fletcher-Flinn, C. M. (1997). Theory of mind and the origin of divergent thinking. *The Journal of Creative Behavior*, *31*(3), 169-179. https://doi.org/10.1002/j.2162-6057.1997.tb00789.x

Taylor, A. H., Cheke, L. G., Waismeyer, A., Meltzoff, A. N., Miller, R., Gopnik, A., Clayton, N. S., & Gray, R. D. (2014). Of babies and birds: complex tool behaviours are not sufficient for the evolution of the ability to create a novel causal intervention. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1787), 20140837. https://doi.org/10.1098/rspb.2014.0837

Taylor, A. H., Elliffe, D., Hunt, G. R., & Gray, R. D. (2010). Complex cognition and behavioural innovation in New Caledonian crows. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1694), 2637-2643. https://doi.org/10.1098/rspb.2010.0285

Taylor, A. H., Hunt, G. R., Holzhaider, J. C., & Gray, R. D. (2007). Spontaneous metatool use by New Caledonian crows. *Current Biology*, *17*, 1504-1507. https://doi.org/10.1016/j.cub.2007.07.057

Tebbich, S., Seed, A. M., Emery, N. J., & Clayton, N. S. (2007). Non-tool-using rooks, Corvus frugilegus, solve the trap-tube problem. *Animal cognition*, *10*(2), 225-231. https://doi.org/10.1007/s10071-006-0061-4

Thom, J. M., & Clayton, N. S. (2013). Re-caching by Western scrub-jays (Aphelocoma californica) cannot be attributed to stress. *PLoS ONE*, *8*(1), e52936. https://doi.org/10.1371/journal.pone.0052936

Tomasello, Call, & Hare. (1998). Five primate species follow the visual gaze of conspecifics. *Animal Behaviour*, *55*(4), 1063-1069. https://doi.org/10.1006/anbe.1997.0636

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson, Organization of memory. Academic Press. https://psycnet.apa.org/record/1973-08477-007

Van der Vaart, E., Verbrugge, R., & Hemelrijk, C. K. (2012). Corvid re-caching without 'theory of mind': A model. *PLoS ONE*, *7*(3), e32904. https://doi.org/10.1371/journal.pone.0032904

Warneken, F., Hare, B., Melis, A. P., Hanus, D., & Tomasello, M. (2007). Spontaneous altruism by chimpanzees and young children. *PLoS Biology*, *5*(7), e184. https://doi.org/10.1371/journal.pbio.0050184

Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science*, *311*(5765), 1301-1303. doi: 10.1126/science.1121448

Weir, A. A. S., Chappell, J., & Kacelnik, A. (2002). Shaping of Hooks in New Caledonian Crows. *Science*, *297*(5583), 981-981. https://doi.org/10.1126/science.1073433

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child Development*, *72*(3), 655-684. https://doi.org/10.1111/1467-8624.00304

Westra, E., & Nagel, J. (2021). Mindreading in conversation. *Cognition*, *210*, 104618. https://doi.org/10.1016/j.cognition.2021.104618

Williamson, T. (2000). The necessary framework of objects. *Topoi*, *19*(2), 201-208. https://doi.org/10.1023/A:1006405915896

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103-128. https://doi.org/10.1016/0010-0277(83)90004-5

Wimmer, H., & Weichbold, V. (1994). Children's theory of mind: Fodor's heuristics examined. *Cognition*, *53*(1), 45-57. https://doi.org/10.1016/0010-0277(94)90076-0

Wood, J. N., Glynn, D. D., Phillips, B. C., & Hauser, M. D. (2007). The perception of rational, goal-directed action in nonhuman primates. *Science*, *317*(5843), 1402-1405. https://doi.org/10.1126/science.1144663

Woodruff, G., & Premack, D. (1979). Intentional communication in the chimpanzee: The development of deception. *Cognition*, *7*(4), 333-362. https://doi.org/10.1016/0010-0277(79)90021-0

Wrangham, R. W. (1980). An ecological model of female-bonded primate groups. *Behaviour*, *75*(3-4), 262-300. https://doi.org/10.1163/156853980X00447

Yamamoto, S., Humle, T., & Tanaka, M. (2012). Chimpanzees' flexible targeted helping based on an understanding of conspecifics' goals. *Proceedings of the National Academy of Sciences*, *109*(9), 3588-3592. https://doi.org/10.1073/pnas.1108517109