Tesis Doctoral

2016

Morality and Mindreading in Nonhuman Animals

Susana Monsó Gil

Licenciada en Filosofía

Programa de Doctorado en Filosofía

Director: Dr. Jesús Zamora Bonilla (UNED)

Codirector: Dr. Mark Rowlands (University of Miami)

Tesis Doctoral

2016

Morality and Mindreading in Nonhuman Animals

Susana Monsó Gil

Licenciada en Filosofía

Programa de Doctorado en Filosofía

Director: Dr. Jesús Zamora Bonilla (UNED)

Codirector: Dr. Mark Rowlands (University of Miami)

Para Mavi y Cote.

ACKNOWLEDGEMENTS

This thesis is the result of four years of hard work, during which I have received help from a great number of individuals. There are three people in particular, however, without whom this dissertation would not have existed as such, and I must begin by thanking them. My first big thank you goes to my advisor, Jesús Zamora, who gave me the opportunity to be a part of his team at UNED and the freedom to pursue this somewhat unorthodox research topic. I am very grateful for his support, his trust in me, and his perpetual good humour. My second big thank you goes to David Teira, Head of the Department of Logic, History, and Philosophy of Science at UNED, who gave me all the strategic and bureaucratic advice that I needed, pushed me to be more ambitious from the very first day, and took great care of me even though I was never his direct responsibility. My last big (immense, gigantic) thank you goes to my external co-advisor, Mark Rowlands, who took on the task of supervising my work out of pure altruism. It's hard to imagine how I could have made it without his thoughtful guidance, his many words of encouragement, and his sincere belief in my project.

I have shared my time at the Department of Logic, History, and Philosophy of Science at UNED with several brilliant graduate students, who have not only become good friends, but also taught me a great deal at our many seminars, reading groups, and informal discussions: Agata Bąk, Álex Díaz, José Ángel Gascón, Javier González de Prado, and Marco Antonio Joven. Most of them have additionally read parts of this dissertation and given me feedback that infallibly helped improve my work.

Over the course of my PhD studies I have been fortunate to enjoy two stays abroad. I first spent three months at the University of Miami, as a result of a short-stay research grant from the Spanish Ministry of Economy and Competitiveness. I would like to thank Mark Rowlands for his help with this, and his willingness to work side by side with me during my time there. Thanks, as well, to Eugene Rosov, Alice Lowe, Sagid Salles, Damien Dupasquier, Shan Haq, and the rest—my adoptive family—, all of whom made me feel at home. I'm also

grateful to the dogs—Chico, Licker, Hennessy, and Unique—who were so much fun to live with and taught me several important lessons on animal morality. During my final year, I also had the chance to spend three weeks in Vienna, as part of a lectureship contract at the Messerli Research Institute. This was an exhilarating and inspiring experience, not only because I learnt so much from teaching my students there, but also because I got the chance to meet and discuss my work with several scholars who have research interests close to mine. I would like to thank Judith Benz-Schwarzburg and Herwig Grimm for this opportunity, as well as Eva Schwarzinger for helping me out with all the necessary paperwork.

During these last four years, I have also attended many conferences, workshops, and seminars where I have sometimes come across inspiration for my own work. These include the many invited talks organised by the Department of Logic, History, and Philosophy of Science at UNED, the ‘Pain in Animals’ conference held at the University of Glasgow in 2013, the 2014 ISAZ meeting at the University of Vienna, and a talk by Françoise Wemelsfelder at the Messerli Research Institute. I have also presented my own work at various venues, where I’ve always received interesting—sometimes crucial—feedback: the ‘What is cognition?’ workshop at the Ruhr-University of Bochum, the 2013 graduate conferences organised by the SEFA and the SIFA, the 3rd Graduate Seminar on Philosophical Research organised by UNED, the 4th Research Workshop on the Philosophy of Biology and Cognitive Science held at the University of Granada, the ‘De-intellectualizing the mind’ workshop at UNED, the philosophy of mind workshop at the Department of Philosophy of the University of Miami, and the 17th edition of the *Taller d’Investigació en Filosofia* held at the Autonomous University of Barcelona. I’m especially grateful for having had the opportunity to present at this last meeting, since the criticism that I received from my commentator, Claudia Compte, as well as from some of the attendants, made me reconsider many aspects of my work and, in the end, served to greatly strengthen my dissertation.

This thesis has further benefited from feedback from various anonymous reviewers for the journals *Argumenta*, *Biology & Philosophy*, *Philosophical Psychology*, and *Mind &*

Language. Several of the arguments I present in this dissertation have also been shaped, in one way or another, by the many conversations that I've had with philosophers and other scholars during these last few years. Special mention is due to Collin Allen, Kristin Andrews, Judith Benz-Schwarzburg, Samuel Camenzind, José Ángel Gascón, Javier González de Prado, Herwig Grimm, Annika Huber, Richard Moore, and Laura Nuño de la Rosa, all of whom provided me with key tips. Thanks, as well, to Mauricio Suárez, for his part played in convincing me to pursue a PhD. And thanks to Cristian Saborido, for constantly reminding me that I was doing fine.

These acknowledgements would not be complete without a mention to my family and friends. I'm lucky enough to have two parents who have unwaveringly supported my choice of career, and made countless sacrifices to ensure I could be whomever I wanted to be. For that, I am immensely grateful. To the rest of my family and friends—they know who they are—, thanks for providing a refuge away from my thesis, and for never doubting that I would make it. Special thanks to Julia Schell and Pablo 'Susu' Suárez, who have understood my academic struggles better than anyone else, and continued to inspire me with their multiple talents and energetic attitude towards life.

My final—and most important—word of thanks goes to my two life companions, who have shared this journey with me and never failed to lift me up when I was falling. My dog, Tote, was the one who sparked my interest in the animal mind, and I'm grateful for all he has taught me ever since—not least that you don't need much intelligence to love and care for others. But my deepest gratitude goes to my partner, Javi, who has been incredibly patient and supportive throughout the many emotional peaks and valleys that come with writing a thesis. Thanks for understanding and, above all, for making me laugh every day.

TABLE OF CONTENTS

RESUMEN

En este trabajo, se defiende la idea de que ser un sujeto moral (es decir, un individuo capaz de actuar con base en motivaciones morales) no requiere la posesión de una teoría de la mente, entendida como la capacidad de atribuir estados mentales a otros. La defensa de esta idea se hace mediante un análisis de la empatía destinado a mostrar cómo se puede conceptualizar como una emoción de carácter moral sin necesidad de apelar a la presencia de una teoría de la mente, lo que implicaría que hay al menos una motivación moral que no requiere esta capacidad. El objetivo último de la tesis es contribuir al debate en torno a la moralidad en animales, mediante la defensa de una concepción desintelectualizada de la moralidad.

La estructura de la tesis es la siguiente. El capítulo 1 contiene la introducción, en la que se especifica la pregunta a responder y su importancia, así como el marco teórico a utilizar, que es el desarrollado por el filósofo Mark Rowlands en su libro *Can Animals Be Moral?* (2012b). Los capítulos 2, 3 y 4 presentan los tres debates en cuya intersección se enmarca esta tesis: el debate sobre la moralidad en animales, el debate sobre la teoría de la mente en animales y el debate acerca de la naturaleza de la empatía. En cada uno de estos capítulos se presenta tanto la evidencia empírica pertinente, como los problemas conceptuales que cada debate acarrea. Se subrayan los conceptos necesarios para el desarrollo de este trabajo y también las tesis comúnmente defendidas que se pretenden refutar. El capítulo 5 contiene una defensa pormenorizada de la tesis principal de este trabajo: la idea de una concepción mínima de la empatía que sea suficiente para contar como emoción moral sin requerir la presencia de una teoría de la mente. Se argumenta que este concepto es lógicamente posible y que numerosos estudios científicos apuntan a su existencia en la naturaleza. En el capítulo 6 se concluye, resumiendo las tesis defendidas y revaluando los tres debates de partida a la luz de estas consideraciones. Asimismo, se esbozan varios caminos por los que la investigación podría continuar a partir de aquí.

1. INTRODUCTION

1.1. THE QUESTION AND WHY IT MATTERS

In this dissertation, I explore the relationship between the capacity to behave morally, and the ability to attribute mental states to others (broadly refered to as 'mindreading'). For the purposes of this investigation, I focus on three questions. The first of these is the following: can we conceive of a form of empathy that can qualify as a moral motivation in the absence of mindreading capacities? The exploration of this first question will be used as a means to answer a second, broader question: does morality require mindreading? The general aim of this dissertation is to defend the idea that this second question should be answered negatively. In discussing these two questions, I aim to contribute to answering a third one: do members of any nonhuman species qualify as moral creatures?

This dissertation is ultimately intended, then, as an addition to the debate on animal morality. I understand the question of animal morality to be distinct from questions related to animal ethics. Animal ethics could be characterised as the discipline aimed at determining which (if any) nonhuman animals are legitimately entitled to moral consideration, as well what the grounds for, and practical consequences of, this entitlement would be. The question of animal morality, in contrast, is one that concerns not our treatment of animals, but the behaviour of animals themselves. Scholars who deal with this question intend to elucidate whether any animals ever behave in ways that can be considered moral—whether members of any nonhuman species should be considered moral agents, or whether morality in this sense is a uniquely human trait.

The debate on animal morality has gathered some momentum in the last few decades, as empirical studies conducted by psychologists and biologists have begun to deliver some unexpectedly suggestive results. However, the debate is still relatively young and, as such, it is often plagued by conceptual confusions, undisclosed assumptions, and dubious presuppositions. Along the following chapters, I will try to pin down some of these problems. My general aim, however, will be to shed some light on this debate by addressing the relation between the property of being a moral individual—an individual who can behave morally—and the possession of mindreading capacities. I want to determine whether being a moral creature necessarily requires the ability to mindread. This is a question that, to the best of my knowledge, has not been thoroughly addressed until now. The main reason behind this is probably the fact that it is commonly assumed to be *obviously true* that one *cannot* be a moral creature without possessing mindreading capacities. Even though this assumption is pervasive, it is rarely made explicit, and even more rarely argued for. I intend to show that the issue may not be as straightforward as it *prima facie* seems, and that minimal forms of morality that do not require the presence of mindreading capacities are a theoretical possibility with an important amount of empirical plausibility. If my arguments in the following chapters are correct, this will make it somewhat more likely that morality may extend beyond the human species.

Why is it important to determine whether animals are moral creatures? Some may be wary of this enterprise, considering perhaps that describing animals' behaviour in moral terms amounts to nothing more than anthropomorphism. This is a legitimate concern. However, following Kristin Andrews and Brian Huss, we can distinguish between *anthropomorphism*, which "involves ascribing a [*sc.* characteristically human] psychological property to an animal when it lacks that property," and *anthropectomy*, which "involves denying a [*sc.* characteristically human] psychological state to an animal who actually has that mental state" (Andrews and Huss 2014, 717). As these authors rightly point out, both anthropomorphism and anthropectomy are "equally errors; they both constitute a failed attempt to describe the world accurately" (Andrews and Huss 2014, 718). If we mistakenly attribute morality to an animal, we would indeed be committing an error, but this error would be no worse than mistakenly denying her morality. There is no reason to fear one mistake over the other. If we are concerned with the scientific endeavour of describing the world, and—as is indeed the case—we find evidence that points *prima facie* to moral capacities in some nonhuman species, we should aim to determine whether morality in fact exists in other species, while being careful to avoid both anthropomorphism and anthropectomy.

Determining whether animals are moral creatures is not only important from a scientific perspective, but also from the point of view of ethics. Morality has long been understood as a feature that distinguishes humanity from the rest of the animal kingdom. It is not uncommon to find authors who use this distinguishing characteristic as a basis for denying moral rights to animals. Henry McCloskey, for instance, argues that "[w]ithout a moral capacity, actually or potentially, there can be … no moral exercise or waiving of a moral right, and hence no moral rights possessed by mammals that lack moral autonomy, actually and potentially" (McCloskey 1987, 79). While this sort of argument can be—indeed, has been—questioned, the fact remains that viewing humans as the only moral creatures may contribute to justifying a view of our species as superior to the rest, and of nature as being somehow at our disposal. This view is exemplified by Tibor Machan, who states that "[n]ormal human life involves moral tasks, and that is why we are more important than other beings in nature," a claim he uses to justify making "the best use of nature for our success in living our lives" (Machan 2002, 10–1). Any research project that explores the continuity between our species and the rest of the animal kingdom has the potential to deliver results that can serve to subvert this view of humanity, and consequently question our widespread exploitation of animals. Moreover, as we shall see, determining that a certain creature qualifies as a moral individual may impose upon us certain specific duties towards her that we would otherwise not have.

Investigating the morality of nonhuman animals may thus serve to bridge the gap between our species and the rest, as well as contribute to shaping the duties we owe nonhuman beings. But this sort of investigation cannot be done without a clear theoretical framework. Before a biologist or a comparative psychologist sets out to determine whether a certain species ever behaves morally, she must first have a clear idea of what ‘behaving morally’ means. Since philosophy—in this particular case, the sub-discipline of meta-ethics—is involved in clarifying issues such as this one, it seems that philosophers may be especially fit for lending a helping hand. This dissertation can be seen as part of a larger project on behalf of philosophers to aid cognitive scientists in the interpretation of available empirical evidence, as well as the design of future studies.

1.2. THEORETICAL FRAMEWORK

This dissertation is meant to be a continuation of the work on animal morality developed by Mark Rowlands, most notably in his book *Can Animals Be Moral?* (Rowlands 2012b). Rowlands defends the possibility that nonhuman animals may sometimes behave on the basis of moral emotions, and count as moral creatures in this sense. His defence of this idea is done on mostly *a priori* grounds, and without focusing on any specific moral emotion. This dissertation is, first of all, an application of Rowlands’ theory to the analysis of a concrete moral emotion—the emotion of empathy—, with the aim of determining what cognitive mechanisms must be in place for an individual to possess this emotion and fulfil the requisites put forward by Rowlands to count as a moral creature. The use of Rowlands’ theory to conceptualise a minimal notion of empathy as a moral emotion will, at the same time, emerge as a promising way of making sense of much empirical data.

My aim in this dissertation is not just to apply Rowlands’ framework to the analysis of a specific moral emotion, but also to develop his theory further. One of the key ideas behind

Rowlands' defence of animal morality is the thesis that morality does not require the presence of a specific kind of higher-order thought: metacognition, or the ability to entertain thoughts about one's own mental states. The other type of higher-order thought is mindreading, or the ability to entertain thoughts about other individuals' mental states. My aim will be to argue that Rowlands' account of moral behaviour can let go of this second type of higher-order thinking as well. As we shall see, the extrapolation of Rowlands' theory to the case of mindreading cannot be done solely on the basis of Rowlands' arguments, for several of the reasons why morality is thought to require mindreading apply exclusively to this type of higher-order thinking, and not to metacognition. I will thus aim to offer further arguments and develop Rowlands' framework in a way that allows us to circumvent these additional problems.

When utilising Rowlands' theory as a framework for my own arguments, I will be leaving aside other philosophical approaches to the defence of animal morality, such as those developed by Steve Sapontzis (1987), Evelyn Pluhar (1995), and David DeGrazia (1996). The choice is not fortuitous. Rather, it has been motivated by a fundamental difference between the approach these authors follow, and the one developed by Rowlands.

Those who deny the possibility of animal morality tend to do so on the basis of certain cognitive deficiencies that are usually attributed to animals and are thought to preclude the possibility of their behaviour being moral. The abilities animals are accused of lacking are things like normative self-government, metacognition, moral judgement, rationality, and so on. Scholars like Sapontzis, Pluhar, and DeGrazia use a common argumentative strategy to defend their claims against this popular position. They first offer empirical examples of moral behaviour in animals, and then provide evidence that animals possess at least a rudimentary form of those cognitive abilities that are thought necessary for morality. The problem with this strategy is that it can easily fall prey to the counter-argument that only those creatures who possess these capacities in full form (i.e. healthy, adult humans) can count as moral. And in fact, these three authors end up concluding that we can ultimately only attribute a primitive form of morality to animals (Sapontzis 1987, 44; Pluhar 1995, 57; DeGrazia 1996, 204).

The strategy followed by Rowlands is different. We can locate it as part of what we've termed the *aponoian* approach to the study of cognition (Rowlands and Monsó, forthcoming). *Aponoia* comes from the Greek words *apó*, "away from," and *noûs*, "intelligence" or "thought." We use it to refer to descriptions of psychological abilities that leave aside all appeal to highly demanding capacities, such as language or metacognition. Rowlands' approach to the topic of animal morality follows the *aponoian* paradigm, because instead of offering evidence that animals possess the cognitive abilities that are usually thought necessary for morality, he focuses on attacking the very idea that such abilities are indeed necessary for morality. He defends a de-intellectualised account of morality itself, and, by way of it, identifies a very clear sense in which animals' behaviour can qualify as fully moral.

The argumentative strategy followed by Rowlands is thus directed at the very foundation of the skeptic's position. As such, it is more promising than the one championed by Sapontzis, Pluhar, and DeGrazia. This is, firstly, because Rowlands' approach is less dependent on a certain interpretation of the empirical data, and thus less likely to stand or fall as research into higher-order capacities in animals continues to deliver its results. But more importantly, in following this approach, Rowlands is avoiding the pervasive tendency amongst scholars to *anthropofabulate*—a term coined by Cameron Buckner, which refers to the generalised habit of "[tying] the competence criteria for cognitive capacities to an exaggerated sense of typical human performance" (Buckner 2013, 853).

Here is an example of *anthropofabulation*. Henry McCloskey argues that, even though some animals may sometimes perform "seeming 'self-sacrificing', 'disinterested', 'benevolent' actions," they cannot be said to possess "moral capacities" because "there is no real evidence of a capacity to make moral judgments, [and] morally to discriminate when self-sacrifice, gratitude, loyalty, benevolence is morally appropriate" (McCloskey 1979, 42). The capacity to engage in moral judgements and determine when a certain attitude is morally appropriate is undoubtedly one that humans possess, and animals probably lack. However, by tying the competence criteria for morality to this capacity, McCloskey is *anthropofabulating*, because it is highly unlikely that the "typical human performance" in moral contexts involves such a demanding use of our intellectual abilities. Moreover, if we witness a human performing a "self-sacrificing," "disinterested," or "benevolent" action, we would have no doubts labelling it moral, regardless of whether the individual had previously engaged in moral reflection. If we are going to investigate whether animals can behave morally, it is only fair that we take as our starting point and competence criterion the lowest common denominator of all (or most) forms of moral behaviour. This, as we shall see, is precisely the strategy followed by Rowlands, and the one that I will adhere to in this dissertation as well.

1.3. STRUCTURE OF THE DISSERTATION

If we were to locate the main thesis of this dissertation in conceptual space, it would be at the intersection of three different debates: the debate on animal morality, the debate on animal mindreading, and the debate on the nature of empathy. The first three chapters of this dissertation will, as a consequence, be devoted to introducing these debates and circling those concepts that I will need for the defence of my thesis, as well as the claims that I will attempt to refute.

Chapter 2 will address the animal morality debate. I will begin by surveying the empirical evidence that suggests that moral behaviour may be present in some animal species. This is of three sorts: experimental, observational, and anecdotal. All together, it suggests that there may be altruistic and cooperative tendencies in some animals, and perhaps a sense of fairness as well. Having gone over the empirical evidence, I will then introduce the theoretical case against animal morality. I will defend the claim that the argument that is commonly used to make this case is problematic, and moreover, that most authors fall prey to a certain fallacy. I will then argue that the best way to sidestep this fallacy is to adopt the theoretical framework offered by Mark Rowlands (2011; 2012a; 2012b). I will sketch the arguments offered by Rowlands to support this theory, and attempt to defend it from some objections that can be found in the literature.

Chapter 3 will focus on the animal mindreading debate. I will begin by offering the reasons why mindreading is commonly thought to be a cognitive prerequisite for morality. Having gone over these reasons, I will offer a brief history of the animal mindreading debate, in order to show why thinking of mindreading as a necessary condition for morality may work against the case for animal morality. Indeed, what little positive results have been obtained in animal mindreading experiments can all be reinterpreted in non-mindreading terms, which some authors take to mean that there is a complete lack of evidence regarding mindreading capacities in other species. For the purposes of this dissertation, I will assume that these authors are right, in order to argue that this would not preclude animals from being moral.

In chapter 4, I will be concerned with the empathy debate. My purpose in this chapter will be twofold. First, I will attempt to give the reader a sense of the intense disagreement that can be found amongst scholars trying to define the notion of empathy. I will show that empathy is used as an *explanans* in two different contexts: debates on social cognition and debates on moral motivation. Within these two contexts, very disparage notions of empathy can be found, and the relation between empathy and mindreading is conceptualised in many varied ways. My second purpose will be to show the reader that there is, nevertheless, one thing about which all scholars who work on empathy agree, and this is the idea that empathy requires mindreading capacities if it is to count as a moral motivation. As a result, the empirical evidence that points to empathy-based moral behaviour in animals is, more often than not, reinterpreted along the lines of a hypothesis that, as I will later show, is based on a false dilemma.

Having introduced these three debates, I will devote chapter 5 to defending the main thesis of this dissertation: the idea that mindreading capacities are not required for the existence of a minimal form of empathy that qualifies as a moral motivation, which, if true, means that there are important grounds for considering that morality does not require mindreading. This thesis will be defended in three steps. First, I will go back to the arguments presented in chapter 3 that defend a necessary connection between mindreading and moral behaviour, and show how we can devise a minimal form of empathy that can serve to refute all these arguments. Then, I will offer a systematic definition of this notion of empathy, and offer a series of reasons for thinking that it may exist in nature. My third and final step will be to argue that there are several reasons that independently suggest that this form of empathy should count as a moral motivation.

Chapter 6 contains the conclusions of this dissertation, where I summarise the points made along the way, and go back to the three debates that constituted my point of departure, in order to see how they remain once my thesis has been defended. After this, I sketch where the research could go from here. In particular, I argue that there are at least four different research projects that are directly opened up by the conclusions of this thesis, and attempt to show how we might go about tackling each of them.

Before concluding this introduction, a final clarification is due. In this dissertation, I follow a naturalistic approach, which means that I will often refer to scientific studies to back up my claims. Unfortunately, an important percentage of the experiments performed on animals that I will refer to are, at the very least, ethically dubious. While I will make a small allusion to this issue in my final chapter, the focus of my dissertation largely prevents me from engaging in constant detours to assess the ethical quality of the experiments I mention. However, given that the topic of this thesis is morality itself, it is my duty to make clear, at this point, that my reference to any specific scientific study does not necessarily imply an endorsement of any kind on my behalf.

2. MORAL ANIMALS?

In this chapter, I will offer an initial approximation to the problem of animal morality. The first aim of this is to delineate a discrepancy that exists between empirical data that points us in one direction—the conclusion that animals may be moral—, and theoretical arguments that point us in the opposite direction—the conclusion that animals cannot be moral—. To illustrate this discrepancy, I will begin by surveying the empirical evidence that has been gathered in the last century and suggests that some nonhuman species may sometimes behave morally. I will then give an overview of the theoretical case against animal morality, which is championed by philosophers and scientists alike, and aims to undermine all this evidence. I will show that, despite the *prima facie* cogency of the arguments that defend the distinctiveness of human morality, a closer examination reveals that they are not only intrinsically problematic, but also based on a fallacious assumption. In the last section, I will argue that the best way to get rid of this pervasive fallacy is to adopt a third moral category into our discourse: the category of moral subjecthood that was introduced by Mark Rowlands (2011; 2012a; 2012b). I will summarise the arguments put forward by Rowlands to defend this moral category, and then deal with some objections that have appeared in the literature. Accordingly, the second aim of this chapter will be to show that the discrepancy illustrated in the first two sections can be best dealt with by shifting the focus in the animal morality debate from the question of whether animals can be moral agents to the question of whether they can be moral subjects.

2.1. THE EVIDENCE

The possibility of animal morality may seem somewhat counterintuitive, especially to those of us who have been trained in philosophy. Indeed, as we shall see in section 2.2, there is an almost unanimous consensus in moral philosophy with respect to the idea that animals cannot behave morally. In the last 70 years, however, there has been an upsurge of empirical evidence that suggests that philosophers may have to revise their intuitions. Before discussing the theoretical case against animal morality, I am going to review the empirical evidence that has been gathered since the beginning of the 20th century and that suggests the presence of some form of morality in some species of animals. This compilation does not intend to be exhaustive—indeed, the amount of suggestive data being gathered is so vast that a comprehensive account would be far too extensive. Additionally, since there is no universally accepted definition of morality, I'm not going to endorse any such definition at this point. What I'm going to describe, then, is a series of paradigmatic examples of what has been presented by some scholars as evidence for animal morality or proto-morality.

The data I will review takes three forms. Firstly, there is *experimental* data. This has been gathered in a laboratory, under controlled conditions, and it is usually thought to be the most compelling form of evidence. Secondy, there is *observational* data. This has been gathered after several weeks, months, or years of systematic and careful observation of animals in the wild or in captivity. And thirdly, there is *anecdotal* data. This takes the form of a one-time accidental observation of a somewhat unusual event in the wild or in a domestic context. It is considered the least convincing form of evidence, and its instances are often dismissed as 'mere' anecdotes (although there are some scholars who disagree, see, e.g.: Mitchell, Thompson, and Miles 1997; Bates and Byrne 2007). For the time being, I will not take a stand on whether any of the examples presented actually constitutes evidence of the existence of morality in animals, for my aim is simply to show that the evidence points *prima facie* to this idea. An evaluation of the data will have to wait until chapters 5 and 6.

2.1.1. Experimental evidence

Experimental evidence in favour of animal morality began to be gathered in the second half of the 20th century and has roughly centred around three clusters.[1] Some experimenters have focused on studying the presence of *cooperative* tendencies in animals. Others have tried to prove that animals have a sense of *fairness*. And lastly, some experiments have tested animals' tendency to respond *altruistically* (e.g. with help or caring behaviour) to others in need. It is worth mentioning at this point that these three clusters are merely one amongst different possible ways of organising the evidence, and insofar as this is an artificial classification, there will be some cases that do not fall so easily into one category or another. This is something that need not concern us at this point.

2.1.1.1. The cooperation cluster

By cooperation I understand an interaction between two or more individuals that benefits all of them, and through which they obtain something better than what they would have obtained on their own—i.e., what is sometimes termed *mutualism*.

Several experimental studies on cooperation have been performed with chimpanzees—

[1] In categorising the experimental evidence into these three clusters, I am roughly following the work by Marc Bekoff and Jessica Pierce: "The heart of wild justice is the suite of moral behaviors that fall into three rough "clusters" (groups of related behaviors that share some family resemblances) that we've used as a fulcrum to organize our material: the *cooperation* cluster (including altruism, reciprocity, honesty, and trust), the *empathy* cluster (including sympathy, compassion, grief, and consolation), and the *justice* cluster (including sharing, equity, fair play, and forgiveness)." (Bekoff and Pierce 2009, xiv)

the species that, together with rats and mice, has been more often tested for the existence of moral tendencies in them. In a pioneer experiment, Meredith Crawford found that pairs of young chimpanzees could be trained to cooperate when the only way of obtaining food was by simultaneously pulling on a rope (Crawford 1937).[2] The age of the experimental subjects seemed to be a determinant factor in their willingness to cooperate in this test. In a much later study conducted by Robert Chalmeau (1994), this author found that older chimpanzees will not cooperate when paired with a dominant chimpanzee that will monopolise the reward. Accordingly, Alicia Melis, Brian Hare, and Michael Tomasello (2006) found that the performance of chimpanzees in a cooperative task depends on the social relationship that obtains between them. Those pairs of chimpanzees that tend to share food outside the test, will be more likely to cooperate spontaneously (i.e. without training) to acquire food in the test itself. In a study performed by Brian Hare *et al.* (2007), the performance of chimpanzees and bonobos in a cooperative food-retrieval task was compared. The authors found that whenever the food reward was easy to share, both species would be equally successful at cooperating. However, whenever the reward was highly monopolisable, bonobos appeared to be much better at cooperating than chimpanzees. The authors attributed this difference in performance to the higher social tolerance in bonobos, who are more prone than chimpanzees to engage in co-feeding.

In a study with capuchin monkeys, Frans de Waal and Michelle Berger (2000) presented pairs of experimental subjects with an apparatus that, when manipulated jointly, gave them access to two cups. On some trials, both cups were baited, while in others, only one of them was. The experimenters found that the monkeys would cooperate to gain access to the cups, and

[2] A video of the experimental results can be downloaded by clicking on “Chimpanzee cooperation task video clip” at the following website: http://www.emory.edu/LIVING_LINKS/media/video.shtml (“Living Links | Video” 2016)

when only one of them was baited, the monkey that had access to it would spontaneously share the reward with its partner. They found that this sharing occurred significantly more often when the apparatus needed a joint effort to be operated, as opposed to when a solo effort sufficed.

Katherine Cronin *et al.* (2005) found that cottontop tamarins will also cooperate to obtain a food reward, in an experiment in which pairs of these monkeys could obtain food when they operated the apparatus jointly. The experimenters also found that cottontop tamarins will manipulate the apparatus significantly less when they are on their own, as compared to when their partner is present. This suggests that they are aware that they need a partner to collaborate with them in order to obtain the reward. A similar result was obtained in a study with Asian elephants (Plotnik et al. 2011).

William Daniel found that rats can also cooperate, and that they can coordinate their actions in order to obtain food and avoid pain. In a series of experiments (Daniel 1942), pairs of rats were placed in a situation in which obtaining food would be accompanied by an electric shock, unless they coordinated themselves so that one pressed a lever that prevented the electric shock while the other was feeding. The results showed that "[t]he rats learned to exchange positions in this situation and at the same time allow sufficient feeding time for each rat to become adequately fed in the course of the experimental session" (Daniel 1942, 368). In a more recent experiment, Claudia Rutte and Michael Taborsky (2007) presented Norway rats with an apparatus that would provide food for a conspecific but not for themselves. They found that the rats would often help their partner obtain food, and that this was more likely to happen if they themselves had received help from a conspecific before.

2.1.1.2. The fairness cluster

In 2003, Sarah Brosnan and Frans de Waal set out to test whether capuchin monkeys had something akin to a sense of fairness. They trained the experimental subjects to exchange a token with the experimenters in order to receive a slice of cucumber. Then, placing them in pairs, they would give one of the subjects a grape (a much more valued reward) in exchange for the token, while the other one received a cucumber slice as a reward for the same task. They found that this second monkey would often reject the reward, sometimes even hurling it out of the cage at the experimenter.[3] The experimenters presented the result as preliminary evidence that capuchin monkeys "dislike inequity" (Brosnan and de Waal 2003, 299).

In 2005, Sarah Brosnan, Hillary Schiff, and Frans de Waal tested chimpanzees using the same experimental paradigm. Chimpanzees also refused to complete the exchange whenever their partner was better rewarded for the task, with the caveat that this aversion to inequity occurred "only in subjects that lived in pairs or in a relatively newly established social group," appearing not to be present in a far older group that possessed a "tightly knit social structure characterized by intense integration and social reciprocity" (Brosnan, Schiff, and de Waal 2005, 257). Chimpanzees did not demonstrate so-called 'advantageous inequity aversion,' that is, they showed no reaction when their own reward was superior to their partner's, but only when they themselves were worse off. This also appeared to be the case with capuchin monkeys:

> In the case of the capuchins, no partner ever shared their grape with a subject who was receiving cucumber ... In fact, in several situations in which the subject rejected the cucumber slice, the partner would finish their grape and then reach through the mesh to take the subject's cucumber and eat it as well! (Brosnan 2006, 176)

However, a more recent experiment by Sarah Brosnan *et al.* (2010) obtained some surprising results. They tested members of this species using this very paradigm and found that "chimpanzees respond behaviourally to receiving more food than a conspecific partner," for

[3] A video of the reaction of one of the experimental subjects can be downloaded by clicking on "Living links video – Capuchin lab" at the following website: http://www.emory.edu/LIVING_LINKS/media/video.shtml ("Living Links | Video" 2016)

those that “received a higher-value grape refused to participate more often when the other chimpanzee received an inferior carrot” (Brosnan et al. 2010, 1236). This suggests the presence of advantageous inequity aversion in chimpanzees.

Friederike Range *et al.* (2009) tested dogs’ aversion to inequity. The subjects were tested in pairs. The experimenter asked each of them, in turn, to give their paw, and in the test condition gave a high-value reward (a sausage) to one of them and a low-value reward (dark bread) to the other. The dogs apparently did not show an aversion to inequity in the quality of the food reward, and would continue to obey the command even if they were receiving a lower-value reward than their partner. However, the experimenters concluded that the dogs did seem to possess some sensitivity to inequity, insofar as they “showed a tendency to a higher refusal rate, a significantly longer hesitation, higher stress levels, and increased looking at the partner when the partner was rewarded and they themselves were not” (Range et al. 2009, 343). Katherine Cronin and Charles Snowdon (2008) tested cottontop tamarins on a cooperative task in which one of the partners was repeatedly rewarded unequally. They also concluded that members of this species show some sensitivity to inequity, as they will cooperate more often when both individuals are rewarded simultaneously or over repeated interactions. Long-tailed macaques have also been found to express inequity aversion when rewarded unequally (Massen et al. 2012).

Lastly, and although it wasn’t proposed as evidence of fairness sensitivity in a nonhuman species, it’s worth mentioning the study conducted by Keith Jensen, Josep Call and Michael Tomasello (2007) to determine the extent to which chimpanzees will engage in vengeful behaviour. The experimental subjects had access to an apparatus that allowed them to deprive a fellow chimpanzee of his food. The experimenters found that the subjects would tend to work the apparatus whenever their conspecific had stolen the food from them, but not when this food had been given to the former directly by the experimenters. The authors concluded that chimpanzees are “vengeful but not spiteful,” for although they engaged in retaliation, they “did not appear to have the decrease in welfare of conspecifics as an end in itself” (Jensen, Call, and

Tomasello 2007, 13048). This could, at least arguably, also be considered as evidence for a sense of fairness in chimpanzees, insofar as the subjects seemed to be discriminating instances when their conspecific deserved a punishment from when she didn't.

2.1.1.3. THE ALTRUISM CLUSTER

The most compelling experimental evidence in favour of animal morality can perhaps be found in the altruism cluster. I will understand an individual to engage in altruistic behaviour whenever she performs an act that alleviates the distress of another individual or that somehow benefits the latter without providing any direct benefit, or at a cost, to the former. The experimental research on the altruistic tendencies in animals is often thought to have begun with a famous experiment performed by Russell Church (1959). He trained rats to press a lever in order to obtain food. In the test condition, an adjacent cage was placed with an electric grid on the bottom, on top of which stood another rat. Whenever the experimental subjects pressed the lever to obtain food, the rat in the adjacent cage would receive an electric shock. Church found that the rats would refrain from pressing the lever upon noticing that it caused a fellow rat to receive an electric shock. A similar experimental result was also obtained in a later study with pigeons (Watanabe and Ono 1986).

Another pioneer experiment in the altruism cluster was the one performed by George Rice and Priscilla Gainer (1962). The experimental subjects—also rats—were presented with either a fellow rat or a Styrofoam block suspended from a harness. The subjects had access to a lever that, when pressed, lowered the stimulus rat, or the block, to the floor of the cage. The experimenters found that the subjects were much more likely to press the lever if what was suspended from the harness was a fellow rat rather than a Styrofoam block. They presented these findings as preliminary evidence that rats were capable of altruism (although they did, cautiously, use this word with scare quotes). Rice and Gainer did, however, point out that the cause of the subjects' behaviour could have been emotional contagion, which would, according to them, rule out the altruistic character of their behaviour. As we will see in later chapters, the dichotomy behind this assertion—either this is emotional contagion or it is moral behaviour—might be a false one.

In another study performed on rats, James Greene (1969) placed the experimental subjects in a cage with two levers, both of which supplied a food pellet, but one of which was much easier to operate than the other. When the rats started showing a preference for the easier lever, the conditions changed, and the preferred lever would supply a food pellet and simultaneously deliver an electric shock to a conspecific in an adjacent cage, while the non-preferred one would also supply a food pellet and terminate the electric shock being delivered to the conspecific. Greene found that those rats that had previously been shocked would prefer to press the second lever and terminate the electric shock that their conspecific was receiving. Valerie Evans and William Braud (1969) also found that rats would refrain from walking down the arm of a T-maze that resulted in an electroshock delivered to a conspecific.

Recently, Inbal Bartal, Jean Decety and Peggy Mason (2011) performed an experiment in which rats were placed into an enclosure that had a restrainer with a fellow rat trapped inside. The free rat could choose to liberate its conspecific. The authors found that, once the free rats learnt how to open the restrainer, they would reliably do so at a short latency whenever there was a fellow rat trapped in it, but not when the restrainer was empty or when there was a toy rat inside. The rats' performance was not due to an expectation of social contact upon opening the restrainer, for when the apparatus was set so that the trapped cagemate would be liberated into a separate compartment, the free rats continued behaving prosocially. When a second restrainer full of chocolate chips was placed into the compartment, the free rat would open both restrainers and share the food with its cagemate.

In an even more recent set of experiments, Nobuya Sato *et al.* (2015), placed a rat into a compartment that was half-way full of water. All the rats used, whose dislike for water had been shown in previous tests, displayed distress signals when in the water. A second rat—the experimental subject—was then placed into an adjacent compartment that had an elevated floor and a door to the first compartment. The walls between both compartments were transparent, so that the subject could see the other rat in the water. The subject had the possibility of opening the door, thus allowing the other rat to exit the water. The experimenters found that the 90% of the subjects displayed door-opening behaviour,[4] at a latency that decreased across trial sessions. When the roles were switched, the rats who had previously been in the water opened the door at an even shorter latency. When the conspecific was not in water, and thus not displaying distress behaviour, only one experimental subject displayed door-opening behaviour. When given a choice between opening a door that gave them access to food and helping their conspecific, the subjects showed no statistically significant difference in the order in which they chose to open the doors. When they opened the food door first, they often allowed their conspecific to share it with them.

In the sixties, Jules Masserman, Stanley Wechkin, and William Terris set out to test whether Church (1959)'s results could be replicated with a different species: rhesus monkeys. In a first series of experiments (Wechkin, Masserman, and Terris 1964), these experimenters trained rhesus monkeys to pull on a certain chain in order to obtain food whenever a red light was turned on, and to pull on a second chain for food whenever a blue light was turned on. They then placed another monkey in an adjacent cage, and this second monkey would receive an electric shock whenever a specific chain out of these two was pulled. The researchers found that six out of ten experimental subjects showed a statistically significant preference for the non-shock chain, and that this result was independent of the dominance hierarchy that obtained between the two individuals. In a follow-up (Masserman, Wechkin, and Terris 1964), they set out to discard further explanations for this finding. In the test condition, ten out of the fifteen monkeys they used showed a statistically significant preference for the non-shock chain and, surprisingly, two further monkeys refrained from pulling either chain (thus resulting in self-

[4] A video of this experimental result can be viewed by clicking on "Supplementary material" at the following link: http://dx.doi.org/10.1007/s10071-015-0872-2

starvation) for five and twelve days, respectively. The authors concluded that "[a] majority of rhesus monkeys will consistently suffer hunger rather than secure food at the expense of electroshock to a conspecific" and that, while these results may be "enhanced by familiarity or previous experience of shock," they are "not significantly related to relative age, size, sex, or dominance" (Masserman, Wechkin, and Terris 1964, 585).

In 2005, Joan Silk *et al.* performed an experiment with two different groups of captive chimpanzees, to see if they would generate benefits for another individual when this entailed no cost to themselves. The experimental subjects were presented with an apparatus that offered them two choices: they could either provide a food reward for themselves and simultaneously for another individual, or they could simply secure the food reward for themselves. They found that, while chimpanzees were obviously motivated to obtain a food reward themselves, the presence of another individual had no impact on the choice they made when handling the apparatus, which the experimenters took to suggest that "chimpanzee behaviour is not motivated by other-regarding preferences" (Silk et al. 2005, 1357).

Felix Warneken and Michael Tomasello (2006) argued that this last experimental paradigm might not have been the best to test chimpanzees' altruism, for members of this species "often compete over food and the drive to acquire food for themselves might preclude their capacity to act on behalf of others" (Warneken and Tomasello 2006, 1301). Instead, they presented young home-raised chimpanzees with a situation in which they could help a familiar human caregiver who was having trouble reaching a certain object. They found that the chimpanzees would spontaneously fetch the object and hand it over to the human, even though they received no reward or praise for it. In a follow-up experiment, Felix Warneken *et al.* (2007) replicated these results with adult semi-free ranging chimpanzees who had not previously interacted with the human they were helping. Again, no rewards were needed to secure their help. The experimenters also found that the chimpanzees who helped would do so even if it entailed a greater cost because it required more physical effort, and that they would also spontaneously offer their help to a conspecific in need.[5]

Despite Silk *et al.*'s (2005) negative results with chimpanzees on the so-called Prosocial Choice paradigm, Victoria Hornet *et al.* (2011) set out to test whether they could obtain positive results with a similar paradigm. This time, the use of complex apparatus was avoided, and instead the experimenters opted for a token-exchange paradigm, something that the chimpanzees being tested were already familiar with. The experimental subjects could exchange two different kinds of tokens with the experimenters: one kind yielded a food reward only for themselves, and the other yielded a food reward for a fellow chimpanzee as well. The authors received positive results this time, as the chimpanzees overwhelmingly favoured the prosocial choice to the selfish one, regardless of the relation of kinship, affiliation, or rank that obtained between both subjects.

Venkat Lakshminarayanan and Laurie Santos (2008) tested capuchin monkeys using the Prosocial Choice paradigm. The experimental subjects could choose between providing a large food reward for themselves and a large reward for their partner, or providing a large reward for themselves and a small reward for their partner. The authors found that the monkeys would reliably choose the option that secured a large reward for their partner. Interestingly, the monkeys would still continue to supply the large reward to their partner when they themselves were receiving a small reward, a result that the authors considered "especially striking given capuchins' known tendency to ... reject rewards that are relatively smaller than those of another monkey" (Lakshminarayanan and Santos 2008, 1000; see section 2.1.1.2). Positive results using

[5] A video of this experimental result can be viewed in the following link: http://commons.wikimedia.org/w/index.php?title=File%3ASpontaneous-Altruism-by-Chimpanzees-and-Young-Children-pbio.0050184.sv005.ogv (*Experiment 3 Chimpanzee Helps in Experimental Condition.* 2007)

this paradigm have also been obtained with common marmosets (Burkart et al. 2007) and cottontop tamarins (Cronin, Schroeder, and Snowdon 2010).

A recent study by James Burkett *et al.* (2016) points towards the presence of altruistic behaviour in prairie voles, a type of rodent. In this experiment, pairs of prairie voles who were usually housed together were separated, and one of them (the demonstrator) was either left in isolation, or subjected to a form of Pavlovian fear conditioning, by presenting her with tones followed by light electric shocks delivered to her feet. The other vole (the observer) was then returned to the cage, and her spontaneous behaviour was observed. It was found that both male and female observers engaged in allogrooming towards the demonstrator for a longer period of time and following a shorter latency when the latter had undergone the shocks, than during the control condition. This allogrooming was found to have an alleviation effect on the demonstrator's distress, as she displayed less anxiety-related behaviours than when she was left alone after the shocks. The demonstrator didn't show an increase in her allogrooming towards the observer compared to baseline conditions, which the authors interpreted as a sign that "the allogrooming response is not a general stress-coping behavior" (Burkett et al. 2016, 376), but rather, that the observers were displaying a form of consolation towards the demonstrators.

2.1.2. OBSERVATIONAL EVIDENCE

The observational evidence is made up of studies performed mainly in the wild, in which data is systematically collected, classified, and analysed after careful long-term observation of animals' spontaneous and natural behaviour. Observational studies may be deemed more ecologically valid than experimental ones, since the settings and situations the animals face are not as artificial, although field experiments are sometimes performed as part of these studies. This type of research may also be preferred because some species or kinds of behaviour cannot be properly studied in the lab, such as the social behaviour of cetaceans (Byrne and Bates 2011).

The biggest disadvantage that comes from observational studies in the wild, however, is that they do not provide the same level of control as experimental studies, which means that certain variables may be unknowingly influencing the animals' behaviour. Nevertheless, some rather interesting evidence for animal morality has been gathered through observational studies. This can, again, be roughly divided into the same three clusters as the experimental evidence.

2.1.2.1. THE COOPERATION CLUSTER

In the seventies and eighties, a set of different observational studies focused on trying to prove the existence of *reciprocal altruism* in animals in the wild. This term was originally coined by Robert Trivers (1971) and it is distinguished from cooperation in that, instead of being a behaviour that benefits all the parties involved more or less simultaneously, it is an exchange of altruistic acts between unrelated individuals and over the course of a certain amount of time. It is often described with the phrase "you scratch my back now, and I'll scratch yours later," and its existence in the wild is controversial and hard to prove, as it implies keeping track of kin relations in a population, in order to rule out *kin altruism* (i.e. the exchange of altruistic acts amongst related individuals). Reciprocal altruism is also distinguished from *generalised reciprocity*, where altruistic acts are administered freely without keeping track of whether the recipient has previously helped or will be able to reciprocate in the future.

One of the first studies to have suggested the presence of reciprocal altruism in a wild population of animals was published by Craig Packer (1977), and it was a study of the coalitions between male olive baboons in the wild. He observed that coalitions would sometimes take place amongst unrelated pairs of baboons in order to attack an opponent, often with the intention of gaining access to an oestrous female. In this sort of cases, only one of the coalition partners would take the oestrous female, while the other (the one whose help had been solicited) would remain fighting the opponent, thus risking serious injury. Packer found that the coalition was more likely to occur when the one that was not benefitted by the coalition had previously been helped in a similar situation by his ally.

Further evidence of reciprocal altruism in other species has been gathered since then. One of the most famous studies along these lines was published by Gerard Wilkinson (1984), who reported the presence of reciprocal altruism amongst vampire bats. These animals will die within three days if they do not find blood to feed on. Wilkinson found that unrelated vampire bats will often engage in a reciprocal exchange of regurgitated blood, and that they are more likely to donate blood to others if they have previously received blood from them. Another famous example of evidence pointing to reciprocal altruism comes from Robert Seyfarth and Dorothy Cheney (1984), who performed field experiments in a population of vervet monkeys to check for the presence of this capacity in this species. They tape-recorded the vocalisations performed by some of these monkeys when soliciting aid from another. They then played these recordings in the presence of some other individual, who had either been previously groomed by the soliciter of aid or not. Cheney and Seyfarth found that members of this species are more likely to help an unrelated individual when that individual has previously groomed them. One further famous study on this topic was published by Frans de Waal and Lesleigh Luttrell (1988), who performed a comparative study of reciprocal tendencies in the behaviour of chimpanzees, rhesus monkeys and stumptail monkeys. They found that these three species are "capable of (1) maintaining mental records of received support, and (2) regulating their own supportive behavior according to a rule of reciprocity"(de Waal and Luttrell 1988, 114).

In a review of anecdotal and observational evidence gathered with respect to dolphins, Richard Connor and Kenneth Norris (1982) claimed that there was enough evidence to attribute reciprocal altruism to different species of dolphins. Interestingly, they also claimed that the gathered data even suggested the presence of generalised reciprocity in these species, for "altruistic acts are dispensed freely and not necessarily to animals that can or will reciprocate" and they "need not necessarily even be confined to the species of the altruistic individual" (Connor and Norris 1982, 370).

Evidence has also been gathered in favour of the hypothesis that some animals engage in reciprocal acts of aggression, in what appears to be something akin to a system of revenge. In their aforementioned comparative study, de Waal and Luttrell (1988) found that chimpanzees were more likely to participate in an attack against a certain individual if that individual had previously taken part in attacks against them. Further, Filippo Aureli *et al.* (1992), who studied the structure of kin-oriented redirection of aggression in Japanese macaques, found that victims of aggression were likely to attack the relatives of their aggressor within one hour of the original conflict. Additionally, the victim's kin would also tend to attack the relatives of the original aggressor.

Lastly, there is also some observational evidence of complex cooperative behaviours amongst animals. In one of the most famous studies in this respect, Christophe Boesch (2002) documented the collaborative hunting strategies of Taï chimpanzees. In up to 77% of their hunts, these chimpanzees will adopt different roles in order to cooperate in hunting down their prey. The different roles include that of *driver* ("a hunter following the prey in a given direction without trying to catch up with them"), *blocker* ("a hunter placing himself in a tree so as to block the progression of the prey"), *chaser* ("a hunter moving quickly after the prey, trying his best to catch up with one"), and *ambusher* ("hunter placing himself in a position where no prey is yet and where he cannot be easily seen, and he will rush toward the prey as soon as it enters his tree") (Boesch 2002, 33). Group members value certain hunting roles more than others, and correspondingly allow those who perform these hunting roles to have more access to meat. Additionally, Boesch observed that cheaters and untalented hunters are penalised, by being allowed less access to meat (Boesch 2002, 40). Cooperative hunting has also been observed in other species, such as African wild dogs (Creel and Creel 1995) and lions (Stander 1992).

2.1.2.2. The fairness cluster

Scientists who study play interactions in different species sometimes claim that their findings suggest the presence of a sense of fairness them. Social play in animals often involves behavioural patterns that are similar to those used in predation or mating. To avoid the misinterpretation of these behavioural patterns during play, animals often use play markers to signal that what is going on is not fighting or mating, but play. Different species of canids, for instance, have been found to use the play bow as a signal (Bekoff 1977). Marc Bekoff himself has argued that in using these play markers and avoiding cheating, different species show a sense of what it is to behave fairly, for "animals often have social expectations when they engage in various sorts of social encounters the violation of which constitutes being treated unfairly because of a lapse in social etiquette" (Bekoff 2001, 82).

Additionally, animals often also engage in *self-handicapping* and *role-reversing* to play 'fairly.' Self-handicapping occurs when an animal does not use its full strength when playing with another individual. Role-reversing takes place when an animal engages in a behavioural pattern that does not correspond to her relative place in the hierarchy. For instance, a dominant playing with a subordinate might voluntarily roll on her back (Bekoff 2001, 84). Hitoshige Hayaki, in an observational study of social play amongst chimpanzees (1985), reported the presence of both self-handicapping and role-reversing. Duncan Watson and David Croft (1996) analysed the play-fighting strategies in a population of captive red-necked wallabies. They reported that 40% of the play fights had at least one role reversal, and that whenever the play partners were of different age classes, the older ones would always engage in self-handicapping and remain very tolerant of their partner's behaviour. Additionally, "[t]hey were never observed to take advantage of their greater size to overpower their partner" (Watson and Croft 1996, 342).

2.1.2.3. The Altruism Cluster

This cluster is mainly composed of studies of post-conflict behaviour in animals. In particular, researchers are on the lookout for two behavioural patterns that serve to alleviate post-conflict distress: *reconciliation* (understood as post-conflict affiliation between opponents) and *consolation* (understood as post-conflict affiliation between one of the opponents and a third party that was not involved in the fight).

Frans de Waal and Angeline van Roosmalen (1979) first confirmed the presence of reconciliation and consolation in a captive group of chimpanzees at the Arnhem Zoo. They found that it was more common for chimpanzees to engage in reconciliation rather than in affiliation with a third party after a fight. Nevertheless, they also found some instances of consolation, which were characterised by the same special behavioural patterns as reconciliation, namely, "'kiss', 'embrace', 'hold-out-hand', 'submissive vocalization' and 'touch'" (de Waal and van Roosmalen 1979, 64). Nobuyuki Kutsukake and Duncan Castles (2004) studied the post-conflict behaviour of wild chimpanzees in the Mahale Mountains and also witnessed reconciliation and consolation, but they could not find any behavioural patterns that were specific to post-conflict affiliation as opposed to general affiliation.

Giada Cordoni *et al.* (2006) compared the post-conflict behaviour in wild and captive gorillas. They observed reconciliation taking place, but not amongst individuals of any sex-age class, but only between adult male-female dyads. They also found that young individuals would often attempt to console the victims of aggressions, even if they were unrelated. Elisabetta Palagi *et al.* (2004) studied a captive population of bonobos and found that consolation and reconciliation were both present, the former being more frequent and generally preceding the latter. The authors hypothesised that, in this species, consolation likely serves as a substitute for reconciliation. In a more recent study performed on a population of bonobos living in semi-natural conditions, Zanna Clay and Frans de Waal (2013) found that bonobos will spontaneously offer consolation to individuals in distress, and that this occurrs across all age and sex classes, although bystanders are more likely to console relatives or individuals with whom they have a close bond. By observing the behaviour of the victim, they saw that self-scratching was considerably decreased after consolation had occurred, which suggests that consolation serves to alleviate the victim's distress. As for reconciliation, they observed that those former opponents that were closely bonded were more likely to reconcile than those that were weakly bonded. In chimpanzees, it has also been found that consolation has a stress-alleviating effect on the victims of aggression, and that it is more likely to take place between individuals that have a valuable relationship (Fraser, Stahl, and Aureli 2008).

Consolation and reconciliation appear not to be restricted to great apes. Evidence has been found of its presence in Tonkean macaques (Elisabetta Palagi et al. 2014), Asian elephants (Plotnik and de Waal 2014), domestic dogs (Cools, Van Hout, and Nelissen 2008), horses (Cozzi et al. 2010), and wolves (Elisabetta Palagi and Cordoni 2009). Furthermore, consolation may even be present in some non-mammalian species. Amanda Seed, Nicola Clayton and Nathan Emery (2007) studied post-conflict behaviour in a captive population of rooks, a species of corvid that is characterised by its high sociality. No evidence was found of reconciliation, that is, former opponents did not engage in post-conflict affiliation. However, the authors found something akin to consolation, as both the victims and the aggressors would engage in affiliation with their social partner at higher levels after a conflict had taken place than during control periods. Additionally, during this third-party affiliation, the rooks were seen displaying a specific type of behaviour that was only witnessed once during the control periods: interlocking the mandibles of their beaks. In another observational study on an avian species, this time ravens, Orlaith Fraser and Thomas Bugnyar (2010) also found an apparent absence of reconciliation, but saw that bystanders would often offer spontaneous consolation to the victims of aggression. Such unsolicited bystander affiliation was more likely to occur whenever the conflict had been intense and the two parties (bystander and victim) shared a valuable relationship.

2.1.3. Anecdotal evidence

Anecdotal evidence is comprised of one-time observations of unusual animal behaviour in the wild or in a domestic context. This sort of evidence, with respect to animal morality, can be divided into two groups that correspond to a grading of the anecdotes themselves in relation to what is usually considered to be their level of "credibility" or "reliability." I will, accordingly, distinguish between grade 1 anecdotal evidence and grade 2 anecdotal evidence.

2.1.3.1. Grade 1 anecdotal evidence

This type of anecdotal evidence is made up of academic papers that relate an unusual event in a serious and scientific manner, where the event described has often been witnessed by more than one individual, and care is taken not to attribute any more cognitive capacities to the animals than what is strictly needed to explain its occurrence. This sort of evidence is dismissed as irrelevant by some scholars, while others consider that we should pay attention to these unusual events, for they may drive future empirical research. Marc Bekoff, for instance, has argued that "the plural of anecdote is data" (Bekoff 2004, 499), and so that, as long as we gather a sufficient number of these stories, we can begin to consider them as evidence.

A large portion of this sort of anecdotal evidence contains reports of animals' reactions to dead or dying individuals. Several of these come from studies on cetaceans. Stan Kuczaj *et al.* (2001) reported the adoption of an orphaned calf on behalf of a 30-year-old captive female bottlenose dolphin. When the calf died four years later and was removed from the pool, the female began jumping vigorously while vocalising loudly, in a display that lasted over ten minutes. She then refused to eat for three days. James Porter (1977) witnessed a group of around thirty false killer whales that remained in very shallow water (where they risk death by stranding) for three days. One of these whales appeared very ill. The rest were seemingly keeping it company. When this individual died, the rest of the group went back into the ocean. Kyum Park *et al.* (2012) witnessed a case of a group of long-beaked dolphins apparently attempting to save a dying conspecific. The group comprised about 12 individuals that were swimming close together and one of them was wriggling about, with seemingly paralysed pectoral flippers. The central dolphins in the group appeared to be helping this last dolphin stay afloat by supporting its body and keeping it in balance. They then assumed what the authors describe as a "raft-like formation" (Park et al. 2012, 4), with the dying dolphin on top, and one other dolphin holding its head up with its beak. When at last the sick dolphin died, five of his fellows remained touching and rubbing its body.

Elephants are also famous for their interesting reactions to both dying and dead conspecifics. Iain Douglas-Hamilton *et al.* (2006) happened to witness the death of an elephant matriarch (Eleanor). The following is their report of the reaction of another, unrelated, elephant (Grace):

> Eleanor was found at 6:14 pm with a swollen trunk which she was dragging on the ground. She had abrasions to an ear and one leg as well as a broken tusk, probably damaged in a previous fall reported by the rangers. She stood still for a while, then took a few slow small steps before falling heavily to the ground at 6:21 pm. Two minutes later, Grace ... rapidly approached her, with tail raised and streaming with temporal gland secretion. She sniffed and touched Eleanor's body with her trunk and foot ... Then she lifted Eleanor with her tusks back on to her feet. Eleanor stood for a short while, but was very shaky. Her back legs began to collapse and she was unable to maintain her upright position ... Grace tried to get Eleanor to walk by pushing her, but Eleanor fell again facing the opposite direction to her first fall. Grace appeared very stressed, vocalizing, and continuing to nudge and push Eleanor with her tusks ... Grace was left by the rest of her family, but continued alone to try and lift Eleanor, with no success. Eleanor was too weak to take advantage of her help. Grace stayed with her for at least another hour as night fell. (Douglas-Hamilton et al. 2006, 94)

Eleanor died later that evening. The authors reported that, for several days, female elephants from five different families approached her carcass and appeared to showed interest in it: sniffing it, touching it, even nudging and rocking it to and fro. Elephants' reactions to the bodies of dead conspecifics was tested with a field experiment performed by Karen McComb, Lucy Baker, and Cynthia Moss (2006). The results showed that elephants will engage in much more interest-related activity when presented with ivory and skulls from dead elephants, than when presented with other objects or skulls from other animals. Additionally, elephants appear to show an equal amount of interest in the skulls of related and unrelated conspecifics.

Further anecdotal evidence refers to instances of care-giving or altruistic behaviour in animals. Thomas Kunz *et al.* (1994), for instance, reported an instance of care-giving behaviour occurring between two female Rodrigues fruit bats, one of which was experiencing difficulties while giving birth. The assisting female was reported to have groomed the anovaginal region of the expectant female, as well as "'tutored' her in a feet-down birthing posture; ... groomed the emerging pup; and ... physically assisted the mother by manoeuvring the pup into a suckling position" (Kunz et al. 1994, 691). Lucy Bates *et al.* (2008) engaged in a systematic analysis of all the anecdotal evidence that has been gathered and points towards the presence of altruistic tendencies in African elephants. They categorised the relevant data into 7 clusters: *coalitions* (two or more individuals working together against another individual), *protection* (protecting a young or injured individual from some danger), *comfort* (protective behaviour directed at an individual that is distressed but not in real danger), *babysitting* (care directed towards a calf that has been separated from its mother), *retrievals* (returning a separated calf to its family), *assisting mobility* (help directed towards an individual that cannot move), and *removing foreign objects* (attempting to remove an object such as a dart or a spear from a fellow individual). They found 17 cases of coalitions, 29 cases of protection (including chasing predators away from newborns, stopping play fights between calves, and preventing calves from moving into dangerous areas), 129 records of comfort, 21 cases of babysitting, 22 cases of retrievals, 28 cases of assisting mobility (all directed at calves), and 3 cases of removing foreign objects. In the last cluster, one of the cases consisted of an adult male removing a tranquilising dart from another male, another consisted of a juvenile investigating a spear protruding from the back of another juvenile, and the last one concerned a mother removing rubbish from her calf's mouth. The authors point out that this last sort of behaviour is especially surprising because there have been no observations of elephants removing vegetation from others' bodies, even though they frequently carry vegetation on them.

Many further anecdotal accounts appear in popular books written by scientists who have spent years of their lives studying animals. Frans de Waal is one such scholar. He has on many occasions given striking accounts of care-giving behaviour in chimpanzees and bonobos. To give but one example, here is his account of an event that he witnessed at Twycross Zoo, starring a bonobo called Kuni:

> One day, Kuni captured a starling. Out of fear that she might molest the stunned bird, which appeared undamaged, the keeper urged the ape to let it go ... Kuni picked up the starling with one hand and climbed to the highest point of the highest tree where she wrapped her legs around the trunk so that she had both hands free to hold the bird. She then carefully unfolded its wings and spread them wide open, one wing in each hand, before throwing the bird as hard she could towards the barrier of the enclosure. Unfortunately, it fell short and landed onto the bank of the moat where Kuni guarded it for a long time against a curious juvenile. (de Waal 1997, 156)

In addition to Frans de Waal, Jane Goodall, who has spent most of her adult life studying chimpanzees, has also given many reports of seemingly moral behaviour in this species. One example is the story of chimpanzee Flint and his mother, Flo:

> Flint ... was eight and a half when old Flo died, and should have been able to look after himself. But, dependent as he was on his mother, it seemed that he had no will to survive without her. ... Flint became increasingly lethargic, refused most food and, with his immune system thus weakened, fell sick. The last time I saw him alive, he was hollow-eyed, gaunt and utterly depressed, huddled in the vegetation close to where Flo had died. Of course, we tried to help him. I had to leave Gombe soon after Flo's death, but one or other of the students or field assistants stayed with Flint each day, keeping him company, tempting him with all kinds of foods. But nothing made up for the loss of Flo. The last short journey he made, pausing to rest every few feet, was to the very place where Flo's body had lain. There he stayed for several hours, sometimes staring into the water. He struggled on a little further, then curled up —

> and never moved again. (Goodall 1990, 197)

A further example of a scholar who is fond of anecdotal accounts of animal behaviour is Marc Bekoff, who has devoted much of his work to the study of social behaviour in canids. He also has an interest in animal morality and has provided us with many suggestive examples, like the following story starring his dog companion Jethro:

> Marc's late dog Jethro once brought home a tiny bunny, whose mother had likely been killed by a mountain lion near Marc's home. Jethro dropped the bunny at the front door and when Marc came to the door he looked up as if to say, "please help." Marc brought the bunny into his house, and put it into a cardboard box with water, carrots, and lettuce. For the next two weeks, Jethro was pinned to the side of the box, refusing to go out for walks and often missing meals. After Marc released the bunny, and for months afterward, Jethro would run to the spot and search around for it. Years later, Jethro saw a bird fly into the window of Marc's car and picked up the stunned ball of feathers and carried it over to Marc, once again seeming to ask for help. Marc placed the bird on the hood of the car, and a few moments later the bird flew off. Jethro watched attentively as it took flight. (Bekoff and Pierce 2009, 108)

Although the popular accounts of moral behaviour in animals offered by authors such as de Waal, Goodall and Bekoff are often related in an informal fashion and with some speculation as to the underlying mechanisms, I still classify them as grade 1 because they come from renowned scientists who are extremely familiar with the species in question, and so less likely to anthropomorphise its members than the average layperson. This, as we shall now see, is not the case for grade 2 anecdotal evidence.

2.1.3.2. Grade 2 anecdotal evidence

This cluster refers to all those popular accounts of animal behaviour that can be found in the media or on the Internet. Reports on this sort of evidence tend to be accompanied by a great deal of speculation as to the underlying cognitive mechanisms, as well as by anthropomorphic projections on the animals involved. This sort of evidence tends to be dismissed by scholars as unreliable, as it is often based on accounts that come from laypeople who are not experienced in the study of the species in question. This inexperience leads to some animal behaviour being grossly misinterpreted. A good example of this sort of misinterpretation is a video[6] that was taken at an aquarium and shows a beluga whale interacting with some children in an apparently teasing manner. The whale lunges herself at the children with an open mouth several times in response to their cries of joy. The whale's behaviour was widely described in the press as 'peek-a-boo' play and the video quickly became viral. Expert on cetacean behaviour Lori Marino, however, pointed out that the whale was in fact behaving aggressively, displaying characteristic threat behaviour towards the children.[7] This shows the extent to which an animal's behaviour can be misinterpreted. The examples I will now refer to should all therefore be taken very cautiously, as they have not been directly collected by scientists.

One of the most popular anecdotes is the story of Binti Jua, a female gorilla at the Brookfield Zoo in Chicago. She became famous when, in August 1996, a three-year-old boy fell into her enclosure, the fall rendering him unconscious. Binti Jua picked him up, cradled him, kept other curious gorillas away, and then finally handed him over to the zookeepers. This incident was widely reported in the press.[8] Both Frans de Waal (e.g.: 2006, 32) and Marc Bekoff (e.g.: Bekoff and Pierce 2009, 1) often refer to this story in their writings on animal morality. Another example they commonly refer to is that of Knuckles, the only chimpanzee in captivity with cerebral palsy. What is remarkable about him is not only the fact that he survives despite

[6] Available at this link: http://www.youtube.com/watch?v=Hh84Oe8JxUQ (Michelle Cotton 2014)

[7] https://www.thedodo.com/why-this-video-of-a-beluga-wha-685343078.html (Messenger 2014)

[8] "15 Years Ago Today: Gorilla Rescues Boy Who Fell In Ape Pit" (2011).

his condition, but also and especially the gentle treatment he gets from his fellow chimpanzees at the Centre for Great Apes in Florida, who apparently understand that he is different (Bekoff and Pierce, 97).[9]

The media often report on further anecdotal accounts of seemingly moral behaviour in animals. Here are some examples that have recently appeared in the press. In May 2013, the press reported the case of a whale killer whale that is missing both the dorsal fin and the right-side pectoral fin. It allegedly survives in the wild despite being unable to hunt for her food, due to the help of its family members, who share their food with her.[10] In May 2014, CCTV cameras in California captured the moment in which a four-year-old was attacked and pulled off his bike by a neighbourhood dog. The boy's family cat appears and chases the dog away. The family later reported to have adopted the cat five years ago, after she followed them home from the park, and that she had formed a strong bond with the boy.[11] In December 2014, a monkey was reported to have saved the life of a fellow monkey that had been electrocuted after walking on the electric wires at an Indian train station. The former monkey appears to be trying to revive her unconscious friend by rubbing, hitting, biting, and dipping him in water. After some minutes, her friend at last shows signs of life.[12] In November 2014, a taxi driver in Mexico filmed a dog trying to revive a fellow dog that had just been hit by a car and was lying dead on the road. The former dog nudges his companion and attempts to pull him away from the traffic. After failing, the dog sits down next to his dead conspecific in what the press described as "hoping apparently that it is going to come round."[13]

[9] ("Center For Great Apes :: Knuckles" 2016)

[10] "Disabled Killer Whale with Missing Fins Survives with the Help of Family Who Hunt for Its Food" (2013).

[11] "Boy Hails Cat Tara That Saved Him from Dog: 'She's a Hero!'" (2014).

[12] "Monkey Saves Dying Friend at Indian Train Station" (2014).

[13] "Heartbreaking Moment a Dog Tries to Pull Its Friend out of the Road" (2014).

With this brief survey, we can see that there is not only a considerable amount of evidence coming from observational and experimental studies, but also a suggestive collection of anecdotes, that point to the existence, in nonhuman animals, of behavioural tendencies that might be labelled 'moral.' While it remains clear that none of the evidence presented in this and the previous sections can constitute conclusive proof of the existence of morality in animals, it does not seem unreasonable to state that all this evidence, taken together, builds a preliminary case for the idea that humans may not be alone in the realm of moral actors. Most philosophers would, however, disagree. Let us now turn to their reasons.

2.2. THE THEORETICAL CASE AGAINST ANIMAL MORALITY

Despite the amount of evidence in favour of animal morality that has been gathered in the last century, most philosophers, and a great majority of scientists too, remain altogether sceptical. In this section, I will review some of the reasons behind this scepticism. I will begin by having a look at the typical argument that is used to defend the claim that animals cannot be moral. I will show that this argument is unconvincing because it relies on unproven empirical claims and an unrealistic view of morality. I will then argue that many scholars find this sort of argument convincing because they have fallen prey to a certain fallacy, which I will call the *amalgamation fallacy*.

2.2.1. A COMMON (AND PROBLEMATIC) ARGUMENT

The defence of the uniqueness of human morality is, more often than not, dispatched with a rather hasty argument that follows the following form:

P_1: Being a moral agent requires X.

P_2: Obviously, animals do not possess X.

C: Therefore, animals cannot be moral agents.

This argument can take two alternative forms, either:

P_1: Being a moral agent requires X.

P_2: X is unattainable without X.*

P_3: Obviously, animals do not possess X.*

C_1: Therefore, they cannot possess X.

C_2: Therefore, they are not moral agents.

Or, keeping in line with the substantive amount of empirical evidence in favour of animal morality:

P_1: Animals may seem to behave morally when they do Y.

P_2: But Y is only an instance of truly moral behaviour when it is accompanied by X.

P_3: Obviously, animals do not possess X.

C: Therefore, animals cannot behave morally.

The X (or X*) in question can refer to different abilities and capacities. Some of the most common are: self-awareness, self-control, reflexive thinking, autonomy, mindreading, and metacognition. In this section, I will show that, despite its general appeal, there are reasons for thinking that this argument, in any of its three forms, is problematic. Let’s begin with an example from a text by Jerome Kagan:

> Human morality is defined by intention, not by behaviour. But because biologists are unable to know the intentions of animals, they classify behaviours that benefit another as altruistic. ... A husband who throws a knife at his wife is called aggressive. But a six-month-old infant who throws a knife at a parent is not regarded as aggressive because we recognize that an intention to hurt is absent. Because animals have no conscious intentions, it is misleading, and theoretically regressive, to describe the animal behaviour with words that have intentionality as a primary feature. (Kagan 2000, 47-8)

The argument being put forward is the following:

P_1: Moral agency requires the possession of conscious intentions.

P_2: Animals have no conscious intentions.

C: Therefore, animals cannot be moral agents.

Kagan is right in that behaving morally undoubtedly requires the presence of some sort of intention behind it. It would seem absurd to say that the elevator doors that slide open automatically when someone's arm is trapped are behaving morally. It is unclear, however, what exactly Kagan has in mind when he denies conscious intentions to animals. If he is thinking of all animals as mere automatons that respond in a reflex-like manner to external stimuli without any intentions behind, then few views on animals are more "theoretically regressive" than this one, reminiscent as it is of behaviouristic tendencies that have long been left behind in comparative psychology. It is unlikely, then, that this is what Kagan is thinking of. What he might have in mind is the idea that animals cannot have conscious intentions *of the right sort.* Although some animals may sometimes behave altruistically, their behaviour, Kagan might say, doesn't count as moral unless they can have the explicit intention of behaving altruistically. This would render his position similar to Michael Bradie's when he wrote that "[a]nimals can act on the basis of altruistic motives but they do not and cannot form intentions to so act" (Bradie 1994, 136; quoted in Waller 1997, 343). This might indeed be true, but, as Bruce Waller has pointed out, the fact that animals cannot conceptualise their own motives as elegantly as a (healthy, well educated) human might, doesn't preclude them from having intentions *of the right sort* (Waller 1997, 343). Their intentions may be altruistic even if they can't recognise them as such because they lack the adequate conceptual repertoire. This, as we shall see, is going to be one of the key ideas behind the model of minimal morality that I will construct in later chapters.

Alternatively, what Kagan might be thinking of when he speaks of "conscious intentions" is the ability to metacognise, that is, to reflect upon one's own mental states. Even though the ability to metacognise is often thought a necessary requisite for the possession of moral agency, section 2.3 below will show how it is possible to construct a cogent account of moral behaviour that does not depend on metacognition. Moreover, if metacognition were in fact a necessary condition for moral agency, the evidence gathered up until now certainly does not warrant such a sweeping generalisation as "animals have no conscious intentions." The question of whether some animals can metacognise is still an ongoing debate with no signs of resolution in the near future (see Hampton 2009; Smith et al. 2009 for reviews). Animals, therefore, may or may not have "conscious intentions" in this sense. There is, as of yet, not enough evidence to exclude all nonhuman species from the realm of moral actors on these grounds.

Christine Korsgaard (2006), who follows in the Kantian tradition, also thinks that the differentiating ability that marks the boundary between moral and non-moral beings is metacognition, insofar as this is what allows us to be conscious of the grounds of our actions as grounds, and thus, according to her, grants us a capacity for normative self-government:

> [T]he capacity for normative self-government and the deeper level of intentional

> control that goes with it is probably unique to human beings. And it is in the proper use of this capacity — the ability to form and act on judgments of what we ought to do — that the essence of morality lies, not in altruism or the pursuit of the greater good. ... We have ideas about what we ought to do and to be like and we are constantly trying to live up to them. Apes do not live in that way. We struggle to be honest and courteous and responsible and brave in circumstances where it is difficult. Even if apes are sometimes courteous, responsible, and brave, it is not because they think they should be. (Korsgaard 2006, 116-7)

The link between metacognition and a "deeper" level of control of our intentions is one that can and has been brought into question (Rowlands 2012b, chapter 7; see also section 2.3 below). Leaving this issue aside for now, Korsgaard is assuming that there is only one possible way of being moral. The claim that the essence of morality lies in normative self-government and not in altruism or in the pursuit of the greater good is something that not all theorists would agree with, especially those who do not feel as comfortable in the Kantian tradition as Korsgaard does. It should, accordingly, not be treated as an unquestionable assumption. Furthermore, even if animals are incapable of explicitly entertaining moral ideals, the evidence suggests that some of them do sometimes live up to them. The monkey that renounced his food for 12 days in a row to avoid shocking a conspecific (Wechkin, Masserman, and Terris 1964) was certainly struggling to be moral in a circumstance in which it was difficult. Perhaps (and most probably) he was incapable of reflecting on the circumstances and of thinking that he "should" refrain from pulling the food chain, but he was still reacting in the appropriate way to a morally salient feature of the situation he was in. Being courteous, responsible, and brave should count for something, even in the absence of moral reflection. After all, as Steven Sapontzis has pointed out, the capacity to reason abstractly and entertain moral principles is a necessary one when it comes to constructing a moral theory, but it may not be necessary when it comes to engaging in "intentional, straightforward moral action" (Sapontzis 1987, 37).

In a slightly different approach, Jerome Kagan asserts:

> The emotion of guilt, which is central to human morality, cannot occur in any primate other than humans because guilt requires the agent to know that a voluntary act has hurt another and the behaviour could have been suppressed. Guilt requires the ability to infer the state of another, to reflect on a past action, to compare the products of that reflection with an acquired standard, to realize that a particular action that violated a standard could have been inhibited, and, finally, to evaluate the self's virtue as a consequence of that violation. Guilt is not a possible state for chimpanzees. Indeed, these animals are unable to make much simpler inferences (Kagan 2000, 48)

Translated into the scheme we provided earlier, this argument becomes:

P_1: Being a moral agent requires the capacity to feel guilt.

P_2: Guilt is unattainable without the ability to:

(1) infer the state of another,

(2) reflect on a past action,

(3) compare the products of that reflection with an acquired standard,

(4) realize that a particular action that violated a standard could have been inhibited, and, finally,

(5) evaluate the self's virtue as a consequence of that violation.

P_3: Chimpanzees are unable to make much simpler inferences than needed to possess guilt.

C_1: Therefore, chimpanzees cannot feel guilt.

C_2: Therefore, chimpanzees cannot be moral agents.

Kagan's argument is unconvincing because its conclusions are based on three premises that are unwarranted. He starts by assuming the centrality of the emotion of guilt. Perhaps he is right and it is, *de facto*, a central emotion in human morality, but is it a necessary one? In a world in which no human were capable of feeling guilt, couldn't we still behave on the basis of other moral motivations and act in morally laden ways? Kagan has not provided any arguments to support his claim of the centrality of guilt, and he has not provided any reasons for thinking that morality could not exist in a guiltless society. Furthermore, the cognitive requirements he claims lie at the basis of guilt are extremely intellectualistic. He seems to have made an effort to describe the most intellectual form of guilt he could come up with and set it as the gold standard. But it is doubtful that guilt only occurs when a person goes through a reflexive process as complex as that. If Kagan genuinely thinks that this is the case, then he should have provided arguments as to why less intellectualised forms of guilt do not suffice to acquire morality. Lastly, Kagan asserts that chimpanzees are incapable of engaging in the requisite reflections because they are incapable of performing much simpler inferences. To back up this claim, he refers to a negative result in a mindreading experiment (Povinelli and Eddy 1996), but as we shall see in the next chapter, the question of whether chimpanzees can attribute mental states to others is still part of an ongoing empirical research programme, and the results of this one experiment should not be taken as conclusive. Additionally, chimpanzees might indeed be incapable of engaging in the requisite reflections, but they might instead possess a less intellectualistic form of guilt. The honest way to proceed to determine whether chimpanzees possess guilt or not would be to review the empirical evidence or devise an adequate experimental paradigm, and then determine whether what was found could count as guilt or not. To begin by constructing an implausibly intellectualistic conception of guilt that is going to preclude all nonhuman animals (and perhaps even some human animals) from possessing it right from the start seems like a plan to preserve human uniqueness at all costs.

Beth Dixon, like Christine Korsgaard, has written extensively against the possibility of animal morality. She states:

> What is missing from de Waal's interpretations of the anecdotes he describes is some reason for thinking that apes have the particular intentional states that capture their appraisals of the morally salient features of another's predicament, and that they are motivated to act to relieve the distress of the person or animal. De Waal describes Binti Jua's action as a "rescue," but he offers us no reason for thinking that this gorilla judged that the child *needed* rescuing. (Dixon 2008, 133)

That is:

P_1: Binti Jua may seem to have behaved morally when she rescued the little boy.

P_2: But her behaviour would only count as a "rescue", and thus, as moral, if she had:

(1) previously appraised the morally relevant features of the situation,

and

(2) been motivated to relieve the distress of the boy.

P_3: Binti Jua, being a gorilla, cannot have performed the required judgements.

C: Therefore, her behaviour was not moral.

The reasoning behind Dixon's assertions will be discussed in more detail in chapters 4 and 5. For now, it suffices to point out that Dixon also seems to be relying on an implausibly intellectualistic conception of morality. If moral behaviour only took place when the relevant "appraisals" had occurred, then this would render much human behaviour amoral. Dixon asserts that these sorts of appraisals are "not susceptible to the kind of quantitative gradualism that evolutionary continuity requires" (Dixon 2008, 141) and, therefore, that the only valid kind to acquire morality is the highly cognitively demanding, linguistically-mediated sort of which only 'normal' adult humans are capable. Again, these requirements seem especially tailored so as to exclude any nonhuman animals from the realm of moral actors and dismiss all the relevant evidence right from the start. Furthermore, the fact that Dixon has not been able to devise a way of gradating the requisite appraisals does not prove that they are, indeed, not susceptible to this form of gradation. There is, in fact, a way around this problem, which we will see in section 2.3, and later in chapters 5 and 6.

This quick survey allows us to see that the problem with this argument, in any of its three forms, is twofold. On the one hand, authors rarely explain in a thorough manner why ability X or property X* are so important for moral agency. Rather, these abilities and properties are simply presupposed to be so, in a way that makes them seemingly designed precisely to preclude all nonhumans from being moral. Additionally, these authors' decide to build their claims upon the most intellectualistic vision of morality possible, so that, to paraphrase Mark Rowlands (2012b, 21), they are not really trying to determine whether animals' motivations and behaviour can be moral, but simply making the rather obvious point that animals cannot be moral in *exactly the same way* a 'normal' adult human at her most lucid and reflexive can.[14] The second problem with this argument is that, even if we grant that X or X* are in fact necessary for moral agency, it is usually simply assumed that all animals lack this property or ability, whereas anyone who has studied the literature on animal cognition will know that experimental and observational studies rarely, if ever, conclusively attribute or deny a certain ability to a nonhuman species.

These two problems—the presupposition of an implausibly intellectualist vision of morality and the reliance on unproven empirical claims—are not, by themselves, enough to debunk the theoretical case against animal morality. However, there is another, much more powerful, problem, and it is the fact that these authors are unknowingly relying on a certain assumption that renders their arguments fallacious. This assumption is what I will call the

[14] "I cannot imagine anyone who would wish to argue that animals are moral agents "in the same way that humans are." The question is not whether animals are moral agents in the same way as humans—normal, adult humans—but, rather, whether animals ... can act morally—act for moral reasons." (Rowlands 2012b, 21)

amalgamation fallacy.[15]

2.2.2. THE AMALGAMATION FALLACY

We have seen that the argument that is commonly used to argue against animal morality is, at worst, problematic, and at best, dependent upon ongoing empirical investigations. In this section, I will show that at the heart of the skepticism towards animal morality lies another problem—a rather crippling confusion, which I will refer to as the *amalgamation fallacy*. This fallacy begins with the conflation of three conceptually and logically independent questions:

> 1. The question of whether animals can behave on the basis of motivations that are moral in character, and consequently engage in behaviours that can be considered moral.[16]
> 2. The question of whether animals can judge their own and others' motivations and behaviour as morally right or wrong.
> 3. The question of whether animals should be held morally responsible for their behaviour, and legitimately praised or blamed for it.

These three questions are, more often than not, confused, amalgamated, and treated as one: the question of whether animals can be moral. In fact, however, each of these questions addresses the presence in animals of a different characteristic, all of which are moral in

[15] I would like to thank Álex Díaz for suggesting this term.

[16] For reasons that shall be made explicit in chapter 5, throughout this section I will assume that behaving on the basis of moral motivations is a sufficient (if not necessary) condition for one's behaviour to be considered moral.

character, but distinct, nevertheless. In discussing the possibility of animal morality, we are thus not dealing with whether animals possess one sole characteristic, but with whether they possess three different ones:

1. The ability to behave morally, i.e., on the basis of moral motivations (M*motivation*).
2. The capacity to judge the morality of one's and others' behaviour (M*judgement*).
3. The quality of being morally responsible for one's behaviour (M*responsibility*).

Throughout the remainder of this chapter, I will refer to these three characteristics (M*motivation*, M*judgement*, and M*responsibility*) as the *moral qualities*. They are, in most cases, tacitly presupposed as inseparable, and accordingly it is argued that a being cannot possess one of them without possessing the other two. However, the moral qualities are conceptually, if perhaps not empirically, separable. To see this, let's imagine the following case. John possesses a severe mental illness that has led him to engage in a series of brutal crimes involving mutilation, torture, and murder. When he is finally hunted down by the police and questioned, he appears to show no remorse, claiming to have performed the crimes because he felt an irrepressible urge to do so, because there is nothing in the world he enjoys more than seeing others suffer. When he is subjected to psychological tests, he appears unable to tell right from wrong, and to understand that what he did was evil. It is obvious that John lacks M*judgement*, for telling right from wrong is one of the abilities that this capacity grants. At the same time, John is, *ex hypothesi*, mentally ill and lacks all form of self-control. It thus seems safe to assume that he is not morally responsible for his behaviour, and cannot be legitimately blamed for it. We are more inclined to say that he *needs help* than that he *deserves a punishment*. He thus lacks M*responsibility*. If these three qualities were inseparable, it would mean that he also lacks M*motivation*—but would we really want to say that his motivations when performing these crimes were not moral? That they weren't evil? That his behaviour, as a result, wasn't morally wrong? On the face of it at least, it seems we can attribute M*motivation* to

John, even in the absence of M*judgement* and M*responsibility*.

M*judgement* also seems separable from M*motivation* and M*responsibility*. Think of the case of Elliot, described by Antonio Damasio (1994, chap. 3). Elliot had suffered from prefrontal brain damage due to the removal of a tumor. Most of his mental abilities had remained intact. Elliot passed all the I.Q. tests performed on him with outstanding grades. More importantly, he seemed perfectly capably of engaging in moral judgements. When presented with ethical dilemmas, he responded like any normal person, following principles that are commonly shared (Damasio 1994, 46). And yet, his decision-making capacities appeared to have been severely compromised. While he was able to "perform moral reasoning at an advanced developmental level" (Damasio 1994, 49), he was "unable to choose effectively, or he might not choose at all, or choose badly" (Damasio 1994, 50). Elliot thus undoubtedly retained M*judgement*, but given that he was unable to use his moral reasoning to guide his decisions, and that his decision-making abilities were impaired to the extent that "his free will had been compromised" (Damasio 1994, 30), he would appear to lack M*motivation*. Additionally, someone whose decision-making abilities are so damaged cannot be legitimately held responsible for what he does, so he would also lack M*responsibility*. The case of Elliot, while extreme and bizarre, illustrates that M*judgement* is also independent of M*motivation* and M*responsibility*.

The case of M*responsibility* is a little trickier, and it is probably from here that all confusion stems. Indeed, it seems that M*responsibility* cannot be separated from M*motivation* and M*judgement*. On the one hand, one cannot be legitimately held morally responsible for one's actions unless one is a being who is first capable of behaving in morally relevant ways, and so M*responsibility* presupposes M*motivation*. Additionally, it seems that unless one is capable of understanding the morality of one's actions, one cannot be legitimately praised or blamed for them. The behaviour of John the killer is morally wrong because it causes suffering, etc., but one of the reasons why he cannot be held morally responsible for his killings is precisely the fact that he doesn't understand that what he's doing is wrong. To be legitimately blamed for his behaviour, he would (at least) need to be capable of understanding its wrongness, and therefore, M*responsibility* also presupposes M*judgement*.[17] M*motivation* and M*judgement* are thus necessary conditions (though not, it seems, sufficient ones) for the possession of M*responsibility*.

The relevant question with respect to animal morality pertains only to M*motivation*. The empirical evidence reviewed in section 2.1. suggests that some animals sometimes behave on the basis of moral motivations (motivations such as empathy, compassion, a sensitivity to fairness, etc.). It is a red herring to argue that the motivations behind these animals' behaviour cannot be moral—that is, that animals cannot possess M*motivation*—because they lack M*judgement* and M*responsibility*, for, as we've seen, M*motivation* may be independent of M*judgement* and M*responsibility*. The question of whether animals can behave morally (that is, on the basis of moral motivations) should, therefore, be answered separately from the question of whether they can judge their own and others' behaviour as moral, or whether they can be legitimately praised or blamed for what they do. This point seems like a fairly straightforward one. And yet, the amalgamation fallacy is constantly committed by scholars who argue against the possibility of animal morality.

A good example of the commission of the amalgamation fallacy is provided by Marc Hauser's (2001) six arguments against moral agency in animals. At the basis of his six arguments is an attempt to show that no matter how compelling the evidence in favour of animal morality is, animals are missing the X that would render their behaviour truly moral. However, by amalgamating the three moral qualities and treating them as one, he is failing to properly address the issue at hand. Let's have a look at his six arguments:

[17] Florian Cova (2013) has positioned himself against this claim. I will address his arguments in chapter 5.

> First, if animals are moral agents, they must place values on the moral emotions. Animals experience a diversity of emotions — fear, anger, surprise. What is controversial, however, is the subjective experience of emotion, what it is like to feel afraid, angry, or surprised. Some authors think that animals experience the moral emotions, feelings such as guilt, shame, and embarrassment. I don't. Nor do I think animals experience empathy. The reason underlying my claim is that I don't believe animals have self-awareness, a sense of self that relies on a richly textured set of beliefs and desires. To experience the moral emotions one must have self-awareness. The mirror test discussed in chapter 5 shows that animals have some sense of their bodies, but says nothing about their thoughts and emotions. Further, the experiments reviewed in chapter 7 suggest that animals lack the capacity to attribute mental states to others. Together, these data argue against the possibility of self-awareness in animals. Even if I am wrong, and we find good evidence for such emotions in animals, we must solve another problem. Among human cultures, the moral emotions are universal. To be human is to have the moral emotions. However, we not only have them, we also place values on them. Guilt and shame are associated with doing something wrong, violating a norm. Showing compassion toward someone who has been injured or deprived of their rights is viewed as a positive action, doing the right thing. Research on animals must therefore establish that animals have moral emotions and view these emotions in a context of right and wrong. (Hauser 2001, 310-1)

Hauser acknowledges that animals can feel fear, anger, and surprise, but he denies them self-awareness, so the notion of self-awareness that he has in mind must be some higher-order capacity, and not simply phenomenal consciousness. He defines self-awareness as "a sense of self that relies on a richly textured set of beliefs and desires", which is extremely vague. With what follows, and the connection he makes to mindreading, it appears that Hauser is thinking of metacognition when he speaks of self-awareness. Even if animals in fact lacked metacognition (which is unclear, as there is evidence that suggests the opposite in some species), this could not exclude them at the outset from the possibility of possessing moral emotions (i.e. M*motivation*), but only of the possibility of reflecting upon them, i.e. a precondition for M*judgement*. Hauser goes on to point out that *even if* animals had moral emotions, it would not matter, for one cannot be a moral agent unless one has the ability to *place values* on these moral emotions, that is, the capacity to reflect upon them and label them as right or wrong. This is an explicit reference to M*judgement*, which is an ability that is relevant only to the possibility of being held responsible for one's actions, for it grants one a certain understanding and the ability to make informed choices. As we've seen, however, it is at least a conceptual possibility that animals may behave on the basis of motivations that are moral in the absence of M*judgement*.

> Second, to act as moral agents, animals must have powerful inhibitory mechanisms that allow them to control their passions and alter their expectations. Under certain circumstances, animals can override their passions, inhibiting powerful desires to mate with a receptive female or feed on a bounty of food. In other situations, animals can inhibit a potential solution to a problem, even when it appears to be the most likely candidate. These two inhibitory mechanisms, however, are limited. (Hauser 2001, 311)

The capacity to inhibit one's impulses adds to the degree of responsibility one has over one's actions (M*responsibility*), but does not speak to its morality (M*motivation*). Consider, again, the case of John, our serial killer. He is incapable of inhibiting his killer impulse. This might make him less responsible for his actions, and so that's why we tend to believe that he should receive psychiatric treatment instead of a punishment. However, his killer impulse is still morally evil, and his killings, as a result, are still morally wrong.

> Third, if animals are moral agents, they must consider the beliefs, desires, and needs of others when planning an action. At present, we have no convincing evidence that

> animals attribute beliefs and desires to others. [...] Similarly, we also lack evidence that animals have access to their own beliefs, reflect on them, and contemplate how particular events in the future might change what they believe. If this lack of evidence correctly reveals a lack of capacity, then animals can certainly cooperate, beat each other to a pulp, and make up after a war. But they can't evaluate whether an act of reciprocation is *fair*, whether killing someone is *wrong*, and whether an act of kindness should be rewarded because it was the *right* thing to do. (Hauser 2001, 312, his emphasis)

Hauser is making two separate points here, both of which have to do with the degree of understanding one possesses over one's actions. First, he states that moral agents must be mindreaders, insofar as the ability to attribute mental states to others is necessary to plan one's actions with an informed understanding of the impact they will have on others' mental lives. He then reiterates the point that animals lack metacognition, and this leads them to lack the capacity to assess their actions as right or wrong. Three points can be made here: first, that there is a considerable amount of evidence that points to the fact that some animals may be capable of at least some forms of mindreading and metacognition—of course, here Hauser would state that only the more intellectually demanding forms of mindreading and metacognition, which surely fall beyond the reach of animals, qualify as the appropriate kind to grant moral agency. Second, that animals may nevertheless have a certain capacity to distinguish (some) fair from (some) unfair situations, and (some) right/wrong behaviours from (some) wrong/right behaviours, as is strongly suggested by the empirical evidence we have reviewed. Of course, Hauser would, again, consider that this "sense" of morality and normativity that some animals may possess is not enough, for they cannot explicitly entertain moral principles or norms. And to this we reiterate our previous point: this only attacks the idea that animals may possess M*judgement* and M*responsibility*, but animals may still be capable of possessing M*motivation*.

> Fourth, if animals are moral agents, they must understand how their actions will impact on the feelings and thoughts of other individuals of their species, and take these consequences into consideration when planning an action. If animals lack the capacity to attribute mental states to others, then they are creatures of the present. This is not to say that they lack the capacity to anticipate or think about the future. They certainly do. What they lack is the capacity to think about how their own actions might change the beliefs of others or cause them pain or happiness, and how risky behaviors such as an attack on a new mother or a competitor might result in their own death, the termination of life and all that it brings. (Hauser 2001, 313)

Hauser introduces here a slight change to his point about lack of mindreading capacities: animals cannot project themselves into the future and fully understand the consequences of their actions. But say John the killer is unable to project himself into the future and grasp the suffering to which he is going to subject his victims' families. Does this make his killer impulse and his murders any less morally wrong? No, although it may ameliorate the extent to which we should hold him responsible for them. Being able to comprehend the consequences of our actions seems, again, linked to M*judgement* and M*responsibility*, but not necessarily to M*motivation*.

> Fifth, moral agents understand the notions of duty and responsibility and use these principles as guiding lights when interacting with others. A variety of animals help each other by cooperating. Male dolphins join together to form coalitions so that they may gain access to females. Chimpanzees often appear to play different roles in their cooperative attempts to hunt prey. Although I tend to think of such cooperation as selfishly motivated, we don't understand what animal cooperators think about their relationship. Do they feel as though it is their duty and responsibility to help those that have helped them in the past? Do they feel as though they have been treated disrespectfully when their partner has reneged on an offer to help? Having a sense of

> duty and responsibility is at the core of being a moral agent, and if animals have no sense of self, I don't believe they should be included as members of this class. (Hauser 2001, 313)

It certainly seems plausible that animals lack a sense of duty and responsibility, but it seems implausible to think that the morality of one's motivations and behaviour hinges exclusively on the existence of a sense of duty and responsibility. Hauser is merely reiterating the same point over and over again: animals cannot be moral agents unless they can reflect upon their actions in a highly intellectualistic manner. But reflection is a way to understanding, and as such, it is at the core of M*judgement*. It need not be necessary for M*motivation*.

> Sixth, if animals are moral agents, they must understand the norms of action and emotion in their society and have the capacity to engage in a revolution when their rights have been violated. We know that in many animal societies, dominant animals attack subordinates if the subordinate has access to a resource that the dominant wants. However, if a subordinate walks in front of a dominant and takes first dibs on an estrous female, is the dominant incensed? If the dominant attacks the subordinate, is this because he thinks that there is a code of conduct, one that dictates who has mating rights, first dibs? Would a subordinate ever think about changing the system, overthrowing the normative responses and feelings that define life in a primate group? Animals certainly respond to violations, to individuals attempting to sneak a mating or a piece of food. However, such responses do not seem to be guided by a sense of what is right, a sense that the violation represents an injustice to the group or the species. Rather, when a dominant attacks a sneaky subordinate, it is because the dominant is selfish, and will do what he can to safeguard the resources. Although subordinates may think that this is unfair, I know of no instance where an animal has attempted to overthrow societal norms. No subordinate has ever built up a coalition of support to derail the system, even though subordinates do overthrow dominants.

> Following the takeover, the rules are the same, even if the new ruler is kinder or more aggressive than the previous ruler. If there is no understanding of normative responses and emotions, and what constitutes a violation, there is no moral system, and there are no moral agents. (Hauser 2001, 313-4)

Hauser is putting forth conditions for moral agency, denying that animals meet them, and, upon realising that, in a sense, they meet them, denying that this sense is enough for moral agency. How can animals "respond to violations" if they have no understanding of "the norms of action and emotion in their society"? Of course they do have a form of understanding—this is precisely what allows them to respond to violations. While they are incapable of meeting in the *agora* to discuss these issues, this does not necessarily mean that they don't have a sense of normativity. Moreover, making the capacity to rebel against the establishment a condition for moral agency seems awkward. It is unclear what the relation between the two is, although I suspect Hauser is merely reiterating what seems to be his only point: that animals cannot reflect upon the notions of right, wrong, fair, unfair, etc., that is, that they cannot possess M*judgement*. Nevertheless, it is unclear that all animals will be incapable of possessing this sort of rebellious capacity. For instance, take the example of Santino, a chimp at Sweden's Furuvik Zoo who routinely throws rocks at visitors and was found to be planning his actions ahead of time:

> Over the years, Santino's operation has become increasingly sophisticated, [...] progressing from simple gathering to fabrication. He has been observed chipping away at the concrete rocks on the island with his hands to sculpt dessert plate-size discs to launch at zoo visitors. In the past decade, zoo workers have witnessed him throwing stones on about 50 separate occasions. Santino has managed to hit a handful

> of gawkers during his assaults but, happily, none have been injured.[18]

In behaving in this sort of way, Santino seems to be rebelling against the establishment. Of course, Hauser can always come up with further conditions that intellectualise truly rebellious actions to the point that excludes Santino from being capable of them. But this would only lead me to conclude that underlying all of Hauser's six conditions appears to be the tacit assumption that unless an individual can act morally *in exactly the same way* a normal adult human can (that is, in a way that grants full-blown M*responsibility*), then that individual cannot be a moral agent.

The six arguments presented by Marc Hauser against the possibility of animal morality are, therefore, unconvincing. We've seen how the amalgamation fallacy lies at the basis of each of them. Additionally, this fallacy undoubtedly plays a major role in the over-intellectualisation of morality that so often underlies the theoretical case against animal morality. This is, in fact, the real problem with Dixon, Kagan, and Korsgaard's arguments. They are basing their position on a conception of morality that amalgamates M*motivation*, M*judgement*, and M*responsibility*. Because they are assuming that M*motivation* cannot exist in the absence of M*responsibility*, they are denying the possession of M*motivation* to any individual that cannot accomplish the highly demanding levels of M*judgement* that are necessary for the presence of M*responsibility*. In order to properly address the question of whether animals can be moral, the amalgamation fallacy needs to be left aside. In the following section, I will defend the best approach for accomplishing this task.

[18] "Planning of the Apes: Zoo Chimp Plots Rock Attacks on Visitors" 2014

2.3. IN DEFENCE OF MORAL SUBJECTHOOD

This section will be a defence of the category of *moral subjecthood*, which was introduced by Mark Rowlands (2011; 2012a; 2012b). This defence will be done in three steps. The first will be to argue that the best way to sidestep the amalgamation fallacy is to incorporate this notion into our conceptual repertoire. The second step will be to sketch the arguments put forward by Rowlands in defence of the coherence of this notion. The third and final step will be to address some objections that have been raised against Rowlands' approach and show how they can be solved.

2.3.1. THE NEED FOR A THIRD MORAL CATEGORY

Debates on the moral status of animals tend to revolve around the question of whether animals are *moral patients*, that is, whether they are legitimate objects of moral consideration whose interests must be taken into account when deciding how to act. The term *moral patient* is usually contrasted to that of *moral agent*, which is thought to apply to individuals that are morally responsible for their actions and can be praised or blamed for them. The distinction between moral patients and moral agents has been a fruitful one in the field of animal ethics, where it was introduced by Tom Regan ([1983] 2004). However, when it comes to the question of animal morality, this distinction has played a major role in securing the appealing nature of the amalgamation fallacy. To see this, let's look at how the distinction between moral agents and moral patients was characterised by Regan:

> In contrast to moral agents, *moral patients* lack the prerequisites that would enable them to control their own behavior in ways that would make them morally accountable for what they do. A moral patient lacks the ability to formulate, let alone bring to bear, moral principles in deliberating about which one among the possible acts it would be right or proper to perform. Moral patients, in a word, cannot do what is right, nor can they do what is wrong. (Regan [1983] 2004, 152; his emphasis)

Moral patients were thus originally characterised as lacking the three moral qualities. Moral patients, according to Regan, cannot "control their own behaviour in ways that would make them morally accountable for what they do," which means they lack M*responsibility*. They also cannot "formulate ... moral principles," and so lack M*judgement*. And, lastly, they "cannot do what is right, nor can they do what is wrong," which means they also lack M*motivation*. Moral agents, in contrast, are understood to possess all three moral qualities. A moral agent can behave on the basis of moral motivations, can judge the morality of hers and others' behaviour, and can be legitimately praised or blamed for her actions.

There is nothing intrinsically wrong with either of these two notions. Indeed, the idea of a 'mere' moral patient, that is, an individual who lacks all three moral qualities but still deserves our moral consideration—because, say, she is sentient—is perfectly coherent. The same applies to the idea of a moral agent, that is, an individual whose highly developed intellectual capacities allow her to possess the three moral qualities. The problem lies in the fact that the lack of a third moral category significantly contributes to the proliferation of the amalgamation fallacy. The three moral qualities are understood, within this framework, as a whole. Either an individual possesses these three qualities, and is thus a moral agent, or she doesn't. If the latter is the case, then she may still be a moral patient, but while moral patients can and should be the objects of our moral consideration, they themselves lack all form of morality. The distinction between moral patients and moral agents thus contributes to the illusion of the inseparability of the three moral qualities.

The distinction proposed by Regan, therefore, leaves no conceptual room for the possibility of an individual that possesses one of the moral qualities, but not the other two. And yet, we've seen that the three qualities are conceptually distinct, and that beings that possess only M*motivation* or only M*judgement* are conceivable. Granted, there is only limited empirical evidence that suggests the existence of individuals who possess M*judgement* while lacking M*motivation* and M*responsibility*, and so it doesn't seem pressing to devise a moral category that applies only to individuals who fulfil this characterisation. However, as we saw in section 2.1, we *do* have a considerable amount of evidence that points to the existence of M*motivation* in some nonhuman species, even in the (apparent) absence of M*judgement* and M*responsibility*. In order to give the question of animal morality the serious consideration it deserves, it is only fair that we incorporate a third moral category into our conceptual repertoire—one that pertains only to the quality of M*motivation*. To this end, I will now understand the question of whether animals are moral, not as the question of whether they are *moral agents*, but as the question of whether they are *moral subjects*. In this I will be following the terminology introduced by Mark Rowlands (2011; 2012a; 2012b).

A *moral subject*, as characterised by Rowlands, is a being whose motivations are moral in character. Moral subjecthood is thus distinguished from moral patienthood in that a moral subject is not *merely* a legitimate object of moral concern, but *also* a subject of moral motivations. At the same time, it is distinguished from moral agency in that moral subjecthood doesn't require the intellectual capacities that grant M*judgement* and M*responsibility*, and so a mere moral subject will only possess the first of the three moral qualities. While for some it may seem obvious that motivations *can* be moral in character even if their possessor lacks M*judgement* and M*responsibility*, the long history behind the amalgamation fallacy may cloud the judgement of many. To avoid the charge of begging the question, let us now turn to the arguments presented by Rowlands in defence of moral subjecthood.

2.3.2. Rowlands' arguments

Mark Rowlands has defended the possibility of animal morality in a variety of texts (Rowlands 2011; 2012a; 2012b; 2013; forthcoming; Rowlands and Monsó, forthcoming). In this section, I will focus on his book *Can Animals Be Moral?* (Rowlands 2012b), as it is here where we can find the most thorough exposition of his view on animal morality. Rowlands follows several different argumentative strategies in this book. To keep things as short as possible, I will only address the ones that are most relevant to the issue at hand: a defence of the coherence of the notion of moral subjecthood.

The main claim that is defended in this book is that there are no conceptual or logical obstacles to the idea that some animals may be *moral subjects*, that is, that they may be capable of behaving on the basis of moral motivations. Obviously, Rowlands does not intend to defend the claim that animals can behave on the basis of *any* kind of moral motivation whatsoever, nor does he intend to argue that *all* the moral motivations that underlie human behaviour may be within the reach of animals. The claim he wishes to defend is one that refers only to a specific kind of moral motivation—*moral emotions*. This is a category that includes "sympathy and compassion, kindness, tolerance, and patience," as well as "their negative counterparts such as anger, indignation, malice, and spite." Additionally, "a sense of what is fair and what is not" is also understood, within this framework, as a moral emotion (Rowlands 2012b, 32). The arguments in this book are aimed at proving that, in order for an individual's behaviour to be motivated by one of these moral emotions, "it is not necessary that she have the ability to reflect on her motives or actions; nor does it require that she be able to explicitly formulate or understand the principles on which she acts" (Rowlands 2012b, 22).

Rowlands begins by establishing what he understands by the term *moral emotion*. Moral emotions, he tells us, are mental states that possess propositional content. This propositional content is of two sorts. On the one hand, moral emotions possess *factual content*. This is what allows an emotion to be distinguished from a mood. When one is in a sad mood, there need not be any particular thing in the world that one's emotion is directed at. For one's sadness to be an emotion, instead of a mood, there has to be a state of affairs in the world that one is sad *about*. On the other hand, emotions also possess *evaluative content*. The state of affairs that the emotion is directed at is evaluated in a certain manner. In the case of moral emotions, this evaluation is moral in character. The state of affairs is judged as morally good/bad/right/wrong/fair/unfair/etc. Rowlands illustrates this with the following example:

> Smith, let us suppose, is indignant that Jones snubbed him. The content of his emotion can, it seems, be specified by the first-person utterance: "Jones snubbed me." However, there is clearly more to Smith's indignation than this. Implicit in the emotion is Smith's idea that Jones was (morally) wrong to snub him. Without this evaluation, Smith's indignation is unexplained. Bound up with Smith's indignation is his belief that he has been hard done by: morally speaking, he deserved better. This combination of the factual and the morally evaluative is a defining feature of what I shall call the morally laden emotions. (Rowlands 2012b, 65)

The key idea in Rowlands' approach is the following: "[e]motions, if they are legitimate, track true evaluative propositions, but they do not require that the subject of an emotion entertain, or even be capable of entertaining, such a proposition" (Rowlands 2012b, 67). This is something that applies both in the case of regular emotions (such as fear, happiness, or sadness), and in the case of morally laden emotions. Thus, in order for Smith to possess indignation, it is not required that he entertain, or be capable of entertaining, the evaluative proposition "It was wrong for Jones to snub me." All that is required is for Smith's emotion to *track* this proposition. This is a technical term that is best understood by looking at the errors that can be implicated in Smith's indignation. These can be of two types, depending on whether they affect the factual content of his indignation, or the evaluative content. Smith's indignation is, firstly, based on the factual proposition "Jones snubbed me." That is, Smith believes that Jones snubbed him. If Jones did not, in fact, snub Smith, but merely failed to notice his presence as they crossed paths in the busy lobby, then Smith's indignation would be *misplaced*, for it would be based on a false belief. The error may instead affect the evaluative content of Smith's emotion, in those cases in which it was not, in fact, wrong for Jones to snub him. Say Jones had overheard Smith criticising him behind his back. Smith would then have no right to be indignant that Jones snubbed him. His indignation would be *misguided*, because it would be based on the false evaluative proposition "It was wrong for Jones to snub me."

What it means for Smith's emotion to track this evaluative proposition is, therefore, that, regardless of whether Smith actually entertains it, there is a truth-preserving relation between Smith's emotion and the proposition itself, such that the non-misguided status of Smith's indignation is what *guarantees its truth*. Indeed, if his indignation is not misguided, then this proposition *must* be true. With this in mind, we are in a position to understand Rowlands' definition of moral emotions:

> An emotion, E, is *morally laden* if and only if (1) it is an emotion in the intentional, content-involving sense, (2) there exists a proposition, *p*, which expresses a moral claim, and (3) if E is not misguided, then *p* is true. (Rowlands 2012b, 68)

Once moral emotions have been defined, Rowlands introduces the notion of a moral subject, which, as we've already seen, is the idea of an individual who, at least sometimes, behaves on the basis of moral motivations. His defence of this notion begins with the exposition of both the *logical* case and the *historical* case against moral subjecthood. The logical case against moral subjecthood is built upon the idea that "to say that X is motivated to act by moral considerations is ... equivalent to the claim that X is responsible for, and so can be praised or blamed for, what it does" (Rowlands 2012b, 88). This, as we saw in the previous section, is one of the confusions behind the amalgamation fallacy. Rowlands (to the best of my knowledge) has been the first philosopher to argue that it is not the same to be motivated to act by *moral* motivations than to be responsible for one's actions. To defend this, he begins by noting that the concepts of evaluation and motivation (generally understood) are distinct:

> The motivation for my behavior is one thing; the evaluation of that behavior is quite another. If hard determinism ... were true, then no one could be morally evaluated for what she does, but it would not follow that she was not the subject of motivational states. (Rowlands 2012b, 90)

A live ant that has been sprayed with oleic acid will be removed from the nest as if it were dead by its fellow ants (Wilson, Durlach, and Roth 1958, 110). The latter are obviously subject to a motivational state—they are motivated, by the oleic acid, to mistakenly remove the live ant from the colony. It does not make sense, however, to hold the ants accountable for this mistake. After all, they behave according to a fixed action pattern that is automatically triggered by chemical signals and, so, could not have done otherwise. The distinction between motivational states and evaluation thus seems like a fairly straightforward one. The problem appears when the motivational states are of a specifically moral character. It is commonly understood that in the absence of an appropriate form of control, and thus in the absence of responsibility, one's motivations cannot be moral. This is, after all, the idea behind the Kantian dictum that *ought implies can*:

> [H]aving control over [one's] motivations is a necessary condition of their possessing normative properties—of their making normative claims on [one]. Ought implies can: without a can there can be no ought. Therefore, without control there can be no normativity. ... Therefore, if an individual is not a moral agent, he cannot be a moral subject either. Being a moral agent is a necessary condition of being a moral subject.

> (Rowlands 2012b, 96)

This is the logical case against the notion of moral subjecthood. Historically, this idea has been shaped by several theories of moral motivation that are dependent upon what Rowlands calls the A-S-C-N-M schema (Rowlands 2012b, 152). This is the pervasive idea that if one does not have ACCESS to one's motivations, then one cannot subject them to SCRUTINY, and that an absence of scrutiny means an absence of CONTROL over one's motivations. Since ought implies can, if one cannot control one's motivations, then they can exert no NORMATIVE grip on us. And lastly, if our motivations are not normative then they cannot be MORAL.

This ASCNM schema is present, firstly, in Aristotle's virtue theory. For an action to be virtuous, according to Aristotle, it is not enough for it to have the right qualities. The agent that performs the action must also fulfil what Rowlands calls the *reflection condition*:

> For any action ϕ, performed by agent A, to be an expression of virtue, V, it is necessary that A (1) be able to understand that ϕ is an instance of V, and (2) perform ϕ because he wishes to be virtuous. (Rowlands 2012b, 102)

In addition to the reflection condition, Aristotle also emphasises the need for phronesis (ϕρόνησις), or practical wisdom. This is what allows an agent to perform an action in the right manner, at the right moment, to the right person, etc., all of which is also necessary for the action to be an expression of a virtue. Both the reflection condition and the phronesis condition are part of the S step in the ASCNM schema. Grace, the elephant who apparently tried to help the dying matriarch Eleanor (see section 2.1.3.1), cannot be said to possess the virtue of compassion because she cannot fulfil either of these two conditions:

> [L]acking in the related abilities of reflection and phronesis, no sense can be given to the claim that Grace is doing what she *ought* to do. Grace is simply the subject of a

> feeling or sentiment on which she cannot reflect, and with respect to which she cannot bring to bear practical reasoning abilities. ... [T]he sentiment has no *normative grip* on Grace. That is, it is not the sort of thing that Grace should either embrace or resist. The sentiment simply does whatever it does to Grace, and she has no say in this matter. Given that Grace has no control over the sentiment and the various behaviors to which it disposes her, no sense can be given to the idea that she *should* endorse or reject the sentiment and resulting behaviors. But moral motivations are essentially things that have a normative grip on their subjects: morally good motivations are ones that should be embraced; morally bad ones should be resisted. Therefore neither Grace's motivation nor her resulting behavior can be regarded as moral. (Rowlands 2012b, 108-9, his emphasis)

Aristotle's virtue theory thus leaves no room for the existence of moral subjects. In his framework, one's motivations cannot be moral unless one has control over them and the ability to reflect upon them. One's actions will not count as virtuous unless one has *chosen* to perform them, and chosen them *for the right reasons*. An action is not virtuous in this scheme unless one has full-blown moral responsibility over it. Moral subjecthood and moral agency, therefore, cannot be separated within this framework.

The same applies to those theories of moral motivation that follow the Kantian tradition, such as Christine Korsgaard's. As we saw in section 2.1.1, Korsgaard argues that metacognition is necessary for morality, insofar as it is what allows us to acquire a form of control over our motivations that renders them moral. Rowlands notes that this is also an expression of the ASCNM schema:

> [In Korsgaard's view], we have a type of control over our motivations ... that no other creature has: we can choose our ends. This is grounded in a uniquely human form of self-consciousness that provides us with "reflective distance" between us and our

> motives that allows us to scrutinize those motives and ask ourselves whether they are the ones we should endorse or reject. This is what makes moral action possible. In short, we have a form of control over our motives that no other creature has; and it is this control that allows us to act morally. Thus, Korsgaard's argument reiterates the tight connection between normativity and control that we ... found to play a central role in the arguments of Aristotle. (Rowlands 2012b, 115)

In the Kant/Korsgaard scheme, then, an individual who cannot access her motivations and subject them to the appropriate form of scrutiny, cannot have the right form of control over such motivations, and consequently, these will be lacking in moral character.

If the ASCNM schema were correct, then not only would *mere* moral subjecthood (i.e. moral subjecthood in the absence of moral agency) be a logical impossibility, it would also be the case that no other animal besides us could be moral, for the required form of scrutiny is extremely intellectually demanding. To defend the notion of moral subjecthood, Rowlands devotes the central chapters of his book to debunking the ASCNM schema. In order to do so, he begins by asking us to imagine an individual called Myshkin who has the following characteristics:

> (M1) (1) Myshkin performs actions that seem to be morally good, and (2) Myshkin's motivation for performing these actions consists in sentiments or emotions that seem to be morally good, but (3) Myshkin is able to subject neither the actions not the sentiments to critical moral scrutiny. (Rowlands 2012b, 125)

Many philosophers, Rowlands notes, would be inclined to say that Myshkin's motivations and resulting actions are not moral, even if Myshkin routinely helps others, rejoicing in their happiness and suffering because of their sadness. In order to strengthen his proposal, Rowlands tweaks his characterisation of Myshkin, so that his actions and motivations no longer merely *seem* to be good, but actually *are* good. If an ideal external spectator, whom Rowlands calls Marlow, were to observe the motivations and actions of Myshkin, he would conclude that they are, in fact, the morally appropriate ones to have in the circumstances. The second characterisation of Myshkin thus goes as follows:

> (M2) (1) Myshkin performs actions that are, in fact, good, and (2) Myshkin's motivation for performing these actions consists in emotions or sentiments that are, in fact, the morally correct ones to have in the circumstances, but (3) Myshkin is able to subject neither the actions nor the sentiments to critical moral scrutiny. (Rowlands 2012b, 128)

This second Myshkin has a sensitivity to some morally salient features of situations, that is, to some of the features that make situations morally good or morally bad, like the happiness and suffering of others. This sensitivity is what triggers in him the morally appropriate emotions. In order to avoid the charge that Myshkin is getting things right *by accident*, Rowlands introduces a third characterisation of him:

> (M3) (1) Myshkin performs actions that are good, and (2) Myshkin's motivation for performing these actions consists in feelings or sentiments that are the morally appropriate ones to have in the circumstances, and (3) Myshkin has these sentiments and so performs these actions in these circumstances because of the operations of his "moral module," which connects perceptions of the morally salient features of a situation with appropriate emotional responses in a reliable way, and (4) Myshkin is unaware of the operations occurring in his "moral module" and so is (5) unable to critically scrutinize the *deliverances* of this module. (Rowlands 2012b, 146)

The term "moral module" is not intended in a psychologically realistic way. The idea is simply to postulate a reliable mechanism that will link the detection of morally relevant features of situations with the adequate emotional and behavioural response. The operations of this reliable mechanism are assumed to be cognitively impenetrable. Thus, although Myshkin 3 is aware of the deliverances of this mechanism (i.e. the emotions that trigger his behaviour), he remains unaware of the processes that underlie them.

Is Myshkin 3 a moral subject? Those who wished to deny it would probably do so on the grounds that "(M3) might mitigate the contingency but it does nothing to alleviate Myshkin's absence of control over what he feels and how he acts" (Rowlands 2012b, 147). Indeed, the defenders of the ASCNM schema would argue that Myshkin's motivations and actions are not moral because he cannot engage in the critical scrutiny that would allow him to control his motivations. "To have control over his motivations," these critics might say, "we need more than their mere nonaccidental generation by a reliable mechanism. Myshkin must himself arrive at these motivations: he must adopt them *for the right reasons*, where these reasons are *available to his conscious scrutiny*" (Rowlands 2012b, 148). Rowlands is going to argue that, despite the intuitive appeal of this criticism, it is based on certain common misconceptions. To do so, he firstly reintroduces Marlow, the ideal moral spectator, who is now going to represent the ASCNM conception of a proper moral subject:

> (M4) (1) Marlow performs actions that are good, and (2) Marlow's motivation for performing these actions consists in feelings or sentiments that are the morally correct ones to have in the circumstances, and (3) Marlow has these sentiments and so performs these actions in these circumstances because of the operations of his "moral module," which connects perceptions of the morally salient features of a situation with appropriate emotional responses in a reliable way, and (4) Marlow has access to the operations occurring in his "moral module" and, therefore, (5) is able to engage in effective critical scrutiny of the deliverances of his module. (Rowlands 2012b, 148)

Marlow is therefore distinct from Myshkin in that he possesses metacognitive abilities. He can access the operations of his “moral module” and can critically scrutinise its deliverances. Myshkin and Marlow both are motivated by morally appropriate emotions and engage in morally correct behaviour, but in contrast to the former, the latter can also *understand* “the reasons why they are the correct things to feel and do in these circumstances” (Rowlands 2012b, 150). This, according to the ASCNM schema, is what allows him to have the sort of control over his motivations that can render them moral. Rowlands is going to challenge this schema by arguing that there are no relevant features that justify granting moral subjecthood to Marlow but not to Myshkin. His focus will be on the S-C step in the schema, and he will defend the claim that “we have no viable understanding of the way in which a subject’s ability to engage in critical scrutiny of its motivations could give that subject control over those motivations” (Rowlands 2012b, 154). His argument will have two parts, to each of which he devotes, in turn, chapters 6 and 7 of his book.

In the first part (Chapter 6), Rowlands focuses on the moral phenomenology of both Marlow and Myshkin, in order to argue that it is not here where we can find the factor that justifies the attribution of moral subjecthood to one but not the other. The essential difference between Marlow and Myshkin, as we’ve seen, is the former’s capacity to metacognise and subject his motivations to critical scrutiny. This, Rowlands states, must be understood as an ability that Marlow is *capable* of exercising, but not as an ability that must be exercised *every* time he is morally motivated to act, for this would be too stringent a requirement and would essentially mean that there are no moral subjects. Marlow’s metacognitive capacity is, in fact, not one sole capacity, but a cluster of different abilities:

> First, there is *recognition*: the ability to recognize a motivation. Then there is *interrogation*: the ability to ask oneself whether this motivation is one that should be embraced or one that should be resisted. Third, there is *judgment*: the ability to assess

> the compatibility between motivation and an antecedently accepted moral principle or proposition. (Rowlands 2012b, 180, his emphasis)

For the ASCNM schema to work, we need a viable explanation of how these scrutinising abilities can give rise to a form of control that Myshkin lacks. Rowlands first considers the possibility that the features of Marlow's scrutinising capacities that underwrite his control may be found in the phenomenological differences between his experience of morality and Myshkin's. Both Marlow and Myshkin experience their motivations, but Marlow can also experience the phenomenology associated with wondering whether they are ones that should be endorsed or rejected, whether they are compatible with certain antecedently accepted moral principles, and so on. It is quite obvious, however, that these phenomenological differences cannot justify the attribution of control to Marlow but not to Myshkin:

> A clear and distinctive phenomenology associated with Marlow's moral deliberations does not show that he has control over those deliberations or the motivations that form their basis. Thus, the hard determinist is, typically, at pains to not deny the phenomenology of freedom. He simply denies that our sense of being free adds up to our actually being free. Similarly, that we seem to have control over our motivations, a seeming that is bound up with the phenomenology of moral deliberation, does not entail that we actually do. (Rowlands 2012b, 164)

In the second part of his argument against the ASCNM schema (Chapter 7), Rowlands defends the idea that there is an inescapable fallacy underlying the historical link between scrutiny and control. This is what he calls the *miracle-of-the-meta*, and it is the common and mistaken assumption that something miraculous happens when we move from the first-order level of our motivations to the higher-order levels of critical scrutiny, so that the latter can escape the problem of control that affects the former. Despite the intuitive appeal of the

ASCNM schema, its reliance on the *miracle-of-the-meta* renders it fallacious, for:

> the very issue of control that arises at the level of motivations is also going to be replicated at the (second, third, fourth, and so on order) level of our evaluation of those motivations. ... If first-order motivations can pull Myshkin this way and that, then second-order evaluations of those motivations can do exactly the same to Marlow. If Myshkin is indeed at the "mercy" of his first-order motivations, as the traditional picture would have us believe, then, logically, Marlow is similarly at the "mercy" of his second-order evaluations of these motivations. In short, second-order evaluation of our first-order motivations cannot lift us above the motivational fray that we think endemic to the first-order motivations, for the simple reason that we can be motivated to evaluate our motivations in one way rather than another. (Rowlands 2012b, 186)

The ASCNM schema, therefore, breaks down at the S-C stage, for scrutiny over one's motivations in no way guarantees control over them. Myshkin and Marlow should both, accordingly, be thought to have as much or as little control over their motivations as the other, because the only form of control that could be thought applicable only to Marlow, that is, control gained through scrutiny, is spurious. This means that the grounds for denying the morality of Myshkin's motivations are fallacious. If we cannot make sense of the denial of the morality of his motivations, we cannot make sense of the denial of his moral subjecthood, either. Marlow and Myshkin should both, accordingly, be considered moral subjects. This should have implications for the animal morality debate:

> [T]he dialectical function of Myshkin is, of course, to go proxy for other animals. To the extent that the case against Myshkin qualifying as a moral subject rests on an unsubstantiated, seemingly mysterious, conception of control, so too does the case

> against (at least some) animals. (Rowlands 2012b, 189)

It seems, therefore, that Myshkin's motivations can indeed be moral in character, and thus that he can be a moral subject, which would mean that there are no reasons to think that (some) animals couldn't be moral subjects as well. However, defenders of the ASCNM schema might not give up that easily and, instead, argue that, in doing away with the link between scrutiny and control, what Rowlands has actually done is not guarantee the morality of Myshkin's motivations, but rather eliminate both the normativity and the morality of Marlow's motivations. The last step in Rowlands' argument will be to reconstruct normativity and morality in a way that makes them no longer dependent upon a problematic notion of control. His account of normativity and morality will be a consequentialist and externalist one, inspired by Julia Driver's work (Driver 2006). It will rely, on the one hand, on objective consequentialism, that is, on the view that good actions are the ones that produce good consequences, independently of the agent's expectations when performing them. On the other hand, it will rely on an externalist view of the moral quality of actions, according to which these will be dependent, at least partly, on features external to the agent. These theories allow us to safeguard the morality of Myshkin's (and, thus, also Marlow's) motivations:

> Myshkin's sensitivity is directed toward objectively good- and bad-making features of situations. It is this that makes his emotions—the experiential expressions of this sensitivity—the sort of things that can be normatively assessed. That is, his emotional responses to situations are ones that can be judged as correct or incorrect. In effect, Myshkin's emotions *track* the good and bad-making features of situations, and this tracking is the sort of thing that can be appropriate or inappropriate—correct or incorrect. Myshkin does not just feel happiness or sadness in the face of a good- or bad-making feature of the situation that presents itself to him: he is, or at least can be, *right* to do. (Rowlands 2012b, 228, his emphasis)

With all this in mind, Rowlands constructs a minimal definition of a moral subject:

> X is a *moral subject* if X possesses (1) a sensitivity to the good- or bad-making features of situations, where (2) this sensitivity can be normatively assessed, and (3) is grounded in the operations of a reliable mechanism (a "moral module"). (Rowlands 2012b, 230, his emphasis)

These conditions are intended as sufficient, but not as necessary. Accordingly, this definition does not exclude the possibility that there may be other ways of being a moral subject. But, if an individual, such as Myshkin, fulfils these three conditions, then that individual is a moral subject, "irrespective of any other perceived lacks or deficiencies" (Rowlands 2012b, 231).

If the arguments presented by Rowlands are correct, it seems that there are no conceptual or logical obstacles against the coherence of the notion of moral subjecthood, which means that it may be the case that some nonhuman species are moral subjects. Of course, the question of *which* species, besides humans, are, in fact, moral subjects is an entirely empirical matter. Despite all the empirical evidence that has been gathered until now, this issue is still far from settled, and it may even be the case that no other species does, in fact, fulfil the requirements for the possession of moral subjecthood. However, it is no longer justified to exclude all nonhuman animals from the realm of moral actors from an *a priori* standpoint, as authors like Beth Dixon, Jerome Kagan, Christine Korsgaard, and Marc Hauser have tried to do.

2.3.3. Can animals *really* be moral? Objections to Rowlands' approach

In the year 2012, the journal *Dilemata* published a paper by Rowlands entitled 'Can animals be moral?' (Rowlands 2012a), which is essentially a summary of the ideas in his 2012b book. The journal featured a series of commentaries to this paper, some of which offered objections to Rowlands' theses. As the remaining chapters will assume the plausibility of Rowlands' proposal, in this section, I will examine some of these objections and attempt to resolve them. Before I do that, however, it's worth pointing out two important differences between how Rowlands' ideas are presented in his 2012a paper and how they are presented in his 2012b book.

The first important difference is found in the way Rowlands chooses to describe his third version of Myshkin. In the *Dilemata* paper, the final characterisation of Myshkin (M3) is presented in the following way:

> (M3) (i) Myshkin performs actions that are good, and (ii) Myshkin's motivation for performing these actions consists in feelings or sentiments that are the morally correct ones to have in the circumstances, and (iii) Myshkin has reasons for having these feelings and performing these actions in these circumstances, but (iv) Myshkin is not aware of these reasons. (Rowlands 2012a, 18, quoted from the original text in English)

In his book, however, it is presented as follows:

> (M3) (1) Myshkin performs actions that are good, and (2) Myshkin's motivation for performing these actions consists in feelings or sentiments that are the morally appropriate to have in the circumstances, and (3) Myshkin has these sentiments and so performs these actions in these circumstances because of the operations of his "moral module," which connects perceptions of the morally salient features of a situation with appropriate emotional responses in a reliable way, and (4) Myshkin is unaware of the operations occurring in his "moral module" and so is (5) unable to critically scrutinize the *deliverances* of this module. (Rowlands 2012b, 146, his

emphasis)

Note how conditions 3 and 4 have changed, and one more condition has been added in the second definition. The focus, in 2012b, is no longer on Myshkin's unconsciousness with respect to the reasons that move him to act, but rather, it is now on his lack of awareness with respect to the operations of his moral module, and his resulting inability to critically scrutinise the deliverances of the module. In what follows, I will assume that the first characterisation of M3 should be interpreted with the second characterisation in mind. Thus, when Rowlands says that Myshkin "has reasons to possess these sentiments and perform these actions," I will assume he is referring to the fact that said sentiments are grounded in the operations of his moral module, and thus are not accidental. When he says that "Myshkin is not conscious of these reasons," I will assume he means he cannot access the operations of his moral module, and is therefore incapable of critically scrutinising its deliverances, resulting in him being unconscious of his reasons for acting *as* reasons. This assumption makes sense within the context of Rowlands' paper, for he states that Myshkin's reasons are "embodied in the non-conscious, sub-personal processing operations that occur in his moral module" (Rowlands 2012a, 19, quoted from the original text in English).

The second important difference pertains to the notion of moral agency and its relation to the ability to scrutinise one's motivations and actions. The definitions of moral agency presented in both works are the same and there are also no relevant differences in the way the critique of the notion of control gained through scrutiny is presented. However, what Rowlands does not do in his 2012a paper, as opposed to his 2012b book, is develop a way to safeguard the notion of moral agency once the idea of control gained through critical scrutiny has been rejected. In his book, Rowlands suggests that the best way of thinking about moral agency is to conceptualise it in terms of understanding, instead of control (Rowlands 2012b, chap. 9). Thus, our unique scrutinising abilities make us different from other animals (at least, assuming they are indeed unable to critically scrutinise their motivations) because they grant us an understanding of our motivations and subsequent actions that animals lack. This means that we can grasp the rightness or wrongness of our actions. This, in turn, is what makes us responsible for them, allowing us to qualify as moral agents. In what follows, I will proceed with this idea in mind, and I will sometimes refer to it to deal with some of the objections, although I am aware that these authors did not have access to the book when they wrote their commentaries.

The following paragraphs contain responses to the objections presented in these commentaries to Rowlands' paper. I will not address every single one of these papers, as some of the authors largely agree with Rowlands' position (e.g. Campos Serena, 2012; Herrera Guevara, 2012; Sánchez Suárez, 2012), or simply disagree with "the way Rowlands chooses to defend it" (Gomila 2012, 69). I will focus on those authors who strongly object to Rowlands' claims, and address a varied range of different objections in order to provide a firm grounding for Rowlands' position, and justify the fact that I will presuppose its cogency in the remaining chapters.

Let us begin with the objections presented by Pablo de Lora (2012). His disagreement stems, at least partly, from a misunderstanding of Rowlands' claims, probably caused by the succinctness with which they are presented in his 2012a paper. This is a problem that will reappear in other commentaries and is especially exemplified by de Lora's first objection:

> [The category of moral subject] does not receive a sufficiently precise characterisation with respect to neither its nature nor the consequences that follow from adopting it. (de Lora 2012, 36, my translation)[19]

The category introduced by Rowlands is novel, and it goes against a few of our

[19] "[U]na categoría, la de "sujeto moral", que, al final, no queda suficientemente precisada ni en cuanto a su naturaleza ni en cuanto a las consecuencias que se siguen de abrazarla."

intuitions in several respects. Thus, it is quite understandable if de Lora feels that it has not been characterised in a sufficiently precise way. But this is hardly a reason to reject Rowlands' notion outright. I will therefore not consider this as a real objection. The next one is a little more interesting:

> The other problematic aspect is the way in which Rowlands prevents this moral sensitivity that some animals would have ... from being a pure "accident." For Rowlands, this is not the case because this moral sensitivity "... is grounded in a mechanism ... that generates certain emotions in given circumstances." I confess I don't understand in what way this circumvents the accidentality of the condition of being moral. (de Lora 2012, 36-7, my translation)[20]

There is a certain ambiguity in the term "accident" as used here, and that may be the reason why de Lora considers that Rowlands hasn't satisfactorily solved this problem. The moral module to which Rowlands refers would be any sort of mechanism that links the perception of certain morally relevant features of situations to a certain emotional reaction. There is a sense in which a mechanism like this would not prevent the being's reaction from being accidental: the being hasn't chosen to have this mechanism, and reacting in this sort of way is simply in her nature—it's merely the result of natural selection. Grace's moral sensitivity is an accident in the sense that it's something that she did not choose to have, but it is *not* an accident in the sense that it will reliably take place, provided certain conditions are met. By

[20] "El otro aspecto problemático es el modo en el que Rowlands evita que la sensibilidad moral de la que estarían dotados algunos animales (los elefantes, pero no sabemos si otras especies también) no sea un puro "accidente". Para Rowlands tal no es el caso porque la sensibilidad moral "… se cimenta en un mecanismo… que genera ciertas emociones en unas circunstancias dadas". Confieso que no entiendo de qué modo así se evita la accidentalidad de la condición de ser moral."

invoking a moral module, Rowlands is trying to circumvent the objection that Grace's reaction to Eleonor's dying was not moral because it was just random, contingent, an accident. The moral module makes Grace's reaction to Eleonor's dying no longer an isolated, inexplicable, random event. Instead, we can now say that, when faced with similar situations, Grace will act in a similar sort of way. And that is why her sensitivity is reliable, instead of being a mere accident.

The rest of de Lora's paper makes a straw man out of Rowlands' argument, as it continues by "granting that he is substantially right," and then going on to assume that "we have no reasons to believe of ourselves that we are so different from elephants, when it comes to moral behaviour." After this, de Lora asks, "but do we really have no reasons?" (de Lora 2012, 37, my translation),[21] to which we could well answer that Rowlands never asked us to believe such a thing. What Rowlands tries to show in his article is that there is one type of moral motivation that animals can share with us. But he would surely grant that there are other moral motivations that we may have and animals cannot. For instance, we can engage in a utilitarian calculus to decide what to do, and this is likely out of the reach of other animals. At the same time, Rowlands is careful to point out that animals may be moral subjects, but that it is doubtful that they will ever acquire moral agency. Therefore, this is also something that distinguishes us from them.

De Lora goes on to "resist this conclusion" (which Rowlands, once again, never asked him to accept) by saying that:

> [W]e are moral agents, and thus different from Grace and all other known elephants, because only we can be conscious of the fact that our control over our moral

[21] "Concedamos que tiene razón en lo sustancial: desde ahora no tenemos razones para creernos tan diferentes, en lo que al comportamiento moral se refiere, de los elefantes. Pero, ¿en serio no las tenemos?"

> motivations may be merely an illusion. (de Lora 2012, 37, my translation)[22]

This, contrary to what de Lora thinks, is perfectly coherent with Rowlands' framework, insofar as, within this framework, moral agency is granted through an understanding of our actions, and not through control. Being conscious of the fact that our control may be an illusion is a form of understanding that distinguishes us from animals and, thus, contributes to our moral agency. But what Rowlands would argue here is that the fact that we possess such a form of higher-order thinking gives us no reason to conclude that animals cannot be motivated to act by moral emotions. Rowlands is not trying to prove that there is no difference between human and animal morality, but merely that there are no conceptual obstacles to the idea that some animals may be, at times, motivated to act by moral emotions.

At the end of his paper, de Lora asserts that Grace is a legitimate object of moral concern regardless of whether or not she qualifies as a moral subject (de Lora 2012, 38). The question of how animals should be treated is, however, the question of whether they are moral patients and what we should do about it, which is not the question Rowlands is trying to answer in this paper. De Lora's assertion is, therefore, irrelevant in this context (especially if we take into account Rowlands' contributions to the animal liberation movement in some of his other texts, e.g. Rowlands 2009). Nevertheless, this concern reappears in several commentaries. Asunción Herrera Guevara (2012, 87, my translation), for instance, warns us of "the risks of treating nonhuman animals with justice if and only if they can be called 'moral subjects'."[23] Kepa Tamames (2012) expresses a similar worry:

[22] "[S]omos agentes morales, y por tanto distintos a Gracia y a todos los elefantes conocidos, porque sólo nosotros podemos ser conscientes de que nuestro control sobre nuestras motivaciones morales puede ser meramente una ilusión."

[23] "[Q]uiero plantear los riesgos de tratar con justicia a los animales no humanos si y sólo si pueden ser llamados "sujetos morales"."

> Unless I am grossly disoriented, I understand that the question that entitles Rowlands' work tries to give reasons for a moral (legal, in short) consideration of [nonhuman animals]. I understand that this means that an intimate connection between the moral quality of a subject and the consideration she deserves (assignation of formal rights) is being presupposed. And this, then, is my question: *Why should we judge this connection as essential—or even important?* (Tamames 2012, 143, his emphasis, my translation)[24]

Tamames and Herrera Guevara are, indeed, "grossly disoriented" when making these remarks, for it is not Rowlands' purpose to address the debate on the moral patienthood of animals in this text. The question of whether animals are moral subjects is independent of the question of whether they are moral patients. This being said, while the question of whether an animal is a moral subject or not is irrelevant to the question of whether she deserves our consideration, it may be relevant when it comes to determining what *sort* of consideration she deserves. I will return to this idea at the end of this section and in chapter 6.

Joana Charterina Villacorta (2012) argues that Rowlands' proposal is "anthropocentric." She begins her argument by summarising Rowlands' ideas, and right from the outset it is clear that she has misunderstood them, for she states:

> Even though [an animal] cannot be conscious of what it is that moves him to act in a certain manner, these reasons exist and are *stored* in his own *moral module* (let's say,

[24] "Salvo que me encuentre gravemente desorientado, entiendo que la pregunta que preside el trabajo de Rowlands trata de aportar razones para una consideración moral (jurídica, en definitiva) de los [animales no humanos]. Con lo que entiendo a mi vez que han de suponerse en íntima conexión la calidad moral de un sujeto y la consideración que esta merece (asignación de derechos formales). Y esta viene a ser mi pregunta: *¿Por qué habría de juzgarse esencial -o siquiera importante- tal conexión?*"

> the set of all these morally adequate and reasoned motivations). (Charterina Villacorta 2012, 43, her emphasis, my translation)[25]

There are a few misunderstandings in this phrase. First of all, the moral motivations that Rowlands is dealing with are moral emotions, and within his framework, a being who possesses a moral emotion must be phenomenally conscious of it, in the sense of there being *something that it is like* to undergo it (see, e.g.: Rowlands 2012b, 146-7). So Grace *is*, in a sense, conscious of what is moving her to try and help Eleonor, insofar as it is an emotion with a specific phenomenal character. But she (presumably) cannot have second-order thoughts about this emotion. So, she cannot reflect upon her motivation. This does not mean that her emotion is stored away in some Freudian subconscious. She is (phenomenally) conscious of her, say, sadness, she just cannot have a thought to the effect that she is sad. Furthermore, her moral module is not a storage unit of all her motivations. Rather, it is a mechanism that ensures that this emotion reliably takes place in certain circumstances. One can see where Charterina Villacorta's confusion might stem from, for Rowlands, as we saw, uses the term "reasons" in this paper to refer to the operations of the moral module, which *are* unconscious. Nevertheless, what the word "reasoned" is doing in that last phrase remains a mystery. Grace's emotion is not the result of a reasoning process, but rather, the result of the operations of her moral module, which may simply be physiological.

Charterina Villacorta goes on to reformulate Rowlands' definition of moral subjecthood:

[25] "Aunque [un animal] no pueda ser consciente de qué es lo que le motiva a actuar de una determinada manera, esas razones existen y están *almacenadas* en un *módulo moral* propio (digamos que el conjunto de esas motivaciones moralmente adecuadas y razonadas)."

> *X is a moral subject if and only if X is, at least sometimes, motivated to act by moral considerations, independently of whether it is conscious or not.* [It is unclear what this last part, "whether it is conscious or not," refers to, but from the following we can infer that Villacorta is referring to the moral considerations.]
>
> Case A: *A nonhuman animal is a moral subject, as she is motivated to act by moral considerations, despite the fact that they are not conscious.*
>
> Case B: *A human being is a moral subject, as she acts on moral considerations, despite the fact that they are not fully conscious.* (Charterina Villacorta 2012, 45, her emphasis, my translation)[26]

It is unclear what exactly Charterina Villacorta means by "conscious," but there is definitely a clear sense in which Grace's motivation *is* conscious, insofar as it is an emotion. What *is* indeed unconscious, as I said earlier, is whatever goes on within her moral module. Perhaps the confusion stems from Rowlands' use of the term "consideration," which suggests some sort of reasoning process. That may be why Charterina Villacorta is focusing more than necessary on the operations of the moral module. But let's concede that the operations of Grace's moral module are fully unconscious, and that the processes that lead to our human moral motivations may often be conscious. The argument that follows next is baffling:

> Rowlands continues to defend a purely *anthropocentric* moral criterion. Despite beginning by assuming each species has its own moral nodule [*sic*], he ends up by

[26] "*X es un sujeto moral si y solamente si X está, por lo menos algunas veces, motivado a actuar por consideraciones morales, independientemente de que sea consciente o no.*

Caso A: *Un animal no humano es un sujeto moral ya que está motivado a actuar por consideraciones morales, a pesar de no ser conscientes.*

Caso B: *Un ser humano es un sujeto moral ya que actúa por consideraciones morales, a pesar de no ser plenamente conscientes.*"

> requiring a human criterion to be able to finally assign a complete morality: its consciousness. With which we just go back to a constant topic in debates on the respect to animal rights: whether they are conscious of their acts, and therefore responsible of them, and consequently subjects of rights (as well as of obligations). (Charterina Villacorta 2012, 45-6, her emphasis, my translation)[27]

Again, there are a lot of confusions here. First of all, Rowlands has never stated that each species has its own moral module. That would be a highly implausible assertion, for "each species" refers to every animal (assuming we're leaving plants, bacteria and fungi aside), from flamingos, through crocodiles, to cockroaches and jellyfish. Rowlands has not asserted that *all* animals are moral. What he has stated is that, in order for a certain behaviour in an animal to qualify as a moral one, some conditions must be met, which include the existence in her of some sort of moral module. It is also confusing why Charterina Villacorta thinks that "consciousness" is a *human* criterion. If Rowlands were requiring that an animal be able to read Spinoza's *Ethics* in order to qualify as moral, then the accusation of anthropocentrism would be fitting, but consciousness is surely something we share with many species. Furthermore, what Rowlands is saying in his paper is that animals can be moral *despite not being conscious* of the processes that lead them to experience certain emotions. Even if they cannot be held responsible for their actions due to their lack of understanding, they can still be moral subjects. Lastly, what he presents in his paper is a defence of moral subjecthood, which has no direct relation to moral patienthood, so he is, in fact, avoiding that "constant topic" that Charterina Villacorta refers to

[27] "Rowlands sigue defendiendo un criterio de moralidad puramente *antropocéntrico*. A pesar de comenzar suponiendo a cada especie su propio nódulo moral, acaba por exigirle un criterio humano para poder asignarle finalmente una moralidad íntegra: su consciencia. Con lo que no hacemos sino retornar a la constante en los debates sobre el respeto a los derechos de los animales: el que sean conscientes de sus actos, por tanto responsables de ellos, y por consiguiente sujetos de derecho (a la par que de obligaciones)."

at the end of that paragraph.

In the final pages of her paper, Charterina Villacorta argues that it is unfair to judge whether animals have morality from the point of view of our species, as it may be the case that each species has its own form of morality. We should thus respect and value each species' actions from the perspective of its own morality. In saying this, she is defending a project similar to that presented by Marc Bekoff and Jessica Pierce (2009).[28] And, in fact, she defines morality in a very similar sort of way as these two authors:[29]

> *X is a moral subject if and only if X acts, at least in the majority of occasions, in accordance with the moral considerations of its species E.*
> *E is a species with moral considerations if and only if E reacts with correcting mechanisms when faced with certain individual acts in its groups.* (Charterina Villacorta 2012, 50, her emphasis, my translation)[30]

It is easy to see that this definition is too lax. In defining morality in such a way, Charterina Villacorta is merely characterising a society with norms, but that does not mean that it will be a society with morality. Furthermore, it is very doubtful that this characterisation would encompass as many species as Charterina Villacorta seems to think. In any case, she presents this definition because she is concerned with "*including within the basic ethical human norms*

[28] "We advocate a species-relative view of morality, recognizing that norms of behavior will vary across species." (Bekoff and Pierce 2009, xii)

[29] "We define morality as a suite of interrelated other-regarding behaviors that cultivate and regulate complex interactions within social groups." (Bekoff and Pierce 2009, 7)

[30] "*X es un sujeto moral si y solamente si X actúa, al menos la mayoría de las ocasiones, por las consideraciones morales de su especie E.*

E es una especie con consideraciones morales si y solamente si E reacciona con mecanismos correctivos ante determinados actos individuales en sus colectivos."

the complete respect to the moral norms of the rest of species" (Charterina Villacorta 2012, 49, her emphasis, my translation),[31] and she is convinced that "[t]he most decisive and relevant thing for animals is the *protection of their interests*" (Charterina Villacorta 2012, 48, her emphasis, my translation),[32] and so, engaging in discussions on whether animals fulfil the requirements for what *we* consider the relevant form of morality is, she contends, a waste of precious time we could be spending helping animals. Rowlands would, of course, agree with the idea that we need to protect animals' interests, but that is a different debate. That is the question of whether animals are moral patients and what we should do about it. What we're dealing with here, once again, is a different question: the question of whether animals can act morally, and morality is understood in Rowlands' framework in a realist way that makes it independent of human judgement. Charterina Villacorta just seems to be missing the point.

Marta Tafalla (2012) presents us with "three observations and a paradox." I will just examine the first two observations, as they are the ones that can be properly considered objections to Rowlands' theses. The first of her observations refers to Rowlands' choice of terminology. She states:

> I think that it would be preferable to substitute the concept of *moral subject* for that of *proto-moral subject*, which would clearly indicate that in him one can find the roots of morality, but not full-blown morality. This would precisely allow us to underline the idea of gradation, which is the key to the evolutionary conception of morality. It would also allow us to better conceptualise the fact that animals are not morally responsible for their acts and we can't judge them for them. (Tafalla 2012,

[31] "Se trata de *incluir dentro de las normas básicas éticas humanas el respeto íntegro a las normas morales de las demás especies*"

[32] "Lo más decisivo y relevante para los animales es la *protección de sus intereses*"

> 55, her emphasis, my translation)[33]

In defending this sort of terminology, it is clear that Tafalla is familiar with the work of authors such as Frans de Waal (e.g. de Waal 1996; 2006), who defends that we can find the roots of morality in certain animals, but that full-blown morality is only present in humans. Rowlands' project is a different one, however. He is trying to conceptualise an entirely new moral category: one that pertains only to moral motivation, and not to moral responsibility. Despite not being responsible for her acts, a moral subject acts on the basis of motivations that fully deserve the label "moral." If he were to use the terminology suggested by Tafalla, this idea would not be appropriately conveyed.

Tafalla's second observation is a defence of the role scrutiny plays in our moral lives. She argues that human beings are not like Marlow, the impartial and ideal spectator that Rowlands' refers to in his paper. Rather, she argues, our scrutinising abilities are imperfect, but they still make a difference:

> We humans *try* to scrutinise our emotions, many times we *want* to resist to them, we *try* to critically analyse our own thoughts and we *try* not to let ourselves be influenced by the circumstances. ... We fumble around and we often fail, but we keep trying because we can differentiate between *is* and *ought*, and it is in that difference where one finds the nucleus of morality. No matter how weak our ability to scrutinise our emotions and resist them is, possessing it radically alters what we are. It is that difference, as small as it may be, what makes us humans able to be morally evaluated

[33] "Considero que sería preferible substituir el concepto de *sujeto moral* por el de *sujeto proto-moral*, que indique claramente que en él se hallan las raíces de la moral, pero no la moral de forma plena. Ello permitiría precisamente subrayar la idea de gradación, que es la clave de la concepción evolucionista de la moral. También permitiría conceptualizar mejor que los animales no son moralmente responsables de sus actos y no podemos juzgarlos por ellos."

> and held responsible for what we do, a responsibility it doesn't make sense to ask of animals. (Tafalla 2012, 58, her emphasis, my translation)[34]

The purpose of Rowlands' discussion of the topic of higher-order scrutiny of our motivations is to show that there is no way in which a process of scrutiny can guarantee control over our motivations. The fact that our scrutinising processes are imperfect makes no difference. It may make a difference phenomenologically, insofar as seeing that we sometimes fall prey to certain motivations and at others we manage to resist them may make us feel like our scrutinising processes have a say in this. But the purpose of Rowlands' discussion is not to argue that we never have control over our motivations, but only that such control cannot be gained via a process of scrutiny. And nothing in Tafalla's argument shows how scrutiny can guarantee control. The fact that we can differentiate *is* from *ought* in no way guarantees that we're able to control our choices. However, Rowlands would agree that our scrutinising abilities do make a difference and indeed make us capable of being morally evaluated and held responsible for our actions. But this is because scrutiny grants us an *understanding* of our motivations and actions that animals lack, and not because it gives us a special type of control.

Ángel Longueira Monelos (2012) presents a very short paper to argue that Rowlands' question is an irrelevant one because it won't make any difference, from a pragmatic point of view, whether animals are moral or not:

[34] "Los seres humanos *intentamos* escudriñar nuestras emociones, muchas veces *queremos* resistirnos a ellas, *procuramos* analizar críticamente nuestros propios pensamientos y *tratamos* de no dejarnos influir por las circunstancias. ... Avanzamos a tientas y muchas veces fracasamos, pero lo intentamos porque podemos diferenciar entre el *ser* y el *deber ser*, y es en esa diferencia donde se halla el núcleo de la moral. Por débil que sea nuestra habilidad de escudriñar nuestras emociones y resistirnos a ellas, poseerla cambia radicalmente lo que somos. Es esa diferencia, por mínima que resulte, la que hace que los humanos podamos ser evaluados moralmente y se nos pueda exigir responsabilidad por lo que hacemos, una responsabilidad que no tiene sentido exigir a los animales."

> [T]he question "Can animals be moral?" lays out a practical pseudo-problem, because it won't matter what our answer is: our behaviour and those of other people will continue to be exactly the same and in any case we will continue talking—or, at least, it will continue to be legitimate from a practical point of view that we talk—as if animals did in fact perform moral actions. (Longueira Monelos 2012, 67, my translation)[35]

There are two responses we can give to this argument. On the one hand, we could say that, if we were to follow this sort of approach, and only pursue those questions whose answers would make a direct and immediate practical difference, then many of our current research projects would have to stop, and this would eventually lead to decay in science and culture. Furthermore, who knows how many great advances in science were inspired or triggered by seemingly 'useless' research projects? It does not seem wise to adopt this purely pragmatic approach. On the other hand, even if we were to adopt it, it seems very unlikely that the question posed by Rowlands is one whose answer will indeed make no practical difference. Even if humans have a tendency to describe other animals' behaviour in anthropomorphic terms, and would continue to do so regardless of what the answer to this question turned out to be, it is also true that an immense majority of people consider this language to be metaphorical and, at the end of the day, view animals as inferior beings whose interests don't count as much as ours (if they do count at all). This has led to all the atrocities that we have performed (and continue to perform) on them. Any project that aims to shift our perspective and raise awareness of the fact that we are not as different from the rest of the animal kingdom as we like to believe has the potential to

[35] "[L]a pregunta "¿Pueden los animales ser morales?" plantea un pseudo-problema práctico, pues no importa cuál sea la respuesta: nuestro comportamiento y el de las demás personas seguirá siendo exactamente el mismo y en cualquier caso seguiremos hablando -o, al menos, seguirá siendo legítimo desde un punto de vista práctico que hablemos- como si los animales realizaran de hecho acciones morales."

make a difference and improve the way we treat animals. For that reason, it deserves the benefit of the doubt and should not be discarded so quickly.

The objections presented by Óscar Horta (2012) are largely determined by the way Rowlands presents his ideas in his 2012a paper, as opposed to how they were later presented in his 2012b book. At the very beginning, Horta quotes Rowlands' definition of 'moral subject' and 'moral agent' in his paper and comments:

> [W]hen we examine the arguments presented by Rowlands, we see that the characterisation of moral agents he uses not always coincides with this [initial definition]. After giving his initial definitions, Rowlands goes on to consider that the difference between moral subjects and moral agents is the fact that the latter *can scrutinise and evaluate their moral motivations* [claim 1]. And at the end he considers that what distinguishes moral agents is that they are *conscious of their reasons for acting* [claim 2]. (Horta 2012, 90, my emphasis, my translation)[36]

Horta considers that claim 1 and claim 2 are incoherent with Rowlands' initial characterisation of moral agency, which was the following:

> X is a moral *agent* if and only if X is (a) morally responsible for, and so can be (b) morally evaluated (praised or blamed, broadly construed) for, its motives and actions. (Rowlands 2012a, 5, his emphasis, quoted from the original text in English)

[36] "[C]uando examinamos la argumentación que presenta Rowlands, vemos que la caracterización que hace de los agentes morales no siempre coincide con esta. Tras dar sus definiciones iniciales, Rowlands pasa a considerar que la diferencia entre sujetos y agentes morales radica en que estos últimos pueden someter a escrutinio y evaluación sus motivaciones morales. Y finalmente considera que lo que distingue a los agentes morales es que son conscientes de sus razones para actuar."

If we take into account chapter 9 of Rowlands' 2012b book, we can see why this definition is, *pace* Horta, coherent with both claim 1 and claim 2. Rowlands establishes that the best way to conceptualise moral agency is to conceive of it as being dependent on an ability to scrutinise and evaluate our motivations (claim 1), insofar as such an ability gives us an understanding of them that allows us to have a certain degree of responsibility over our motives and actions. This form of understanding is enabled because we can metacognise and thus are, to a certain extent at least, conscious of the reasons that move us to act in a certain way (claim 2). There is, therefore, no incoherence involved in taking together these three claims.

Another objection presented by Horta turns on the way Rowlands chooses to characterise Myshkin (M3) in this paper:

> Rowlands concludes that the only thing that distinguishes Marlow from Myshkin is [the fact that] Marlow would be conscious of his reasons to act.
>
> There are reasons to think, however, that the difference between Myshkin and Marlow is bigger than that. Let's consider the case of someone who also possesses moral subjecthood, Méndez. Let's suppose Méndez acts in the same way as Myshkin and Marlow, and that in the same way as Marlow, he is conscious of his reasons for acting. Let's suppose, however, that Méndez is incapable of performing any critical evaluation of his motivations to act. He is conscious of them, but is totally carried away by them without being able to perform any sort of reflection with regards to them.
>
> Well, I think that Méndez is conceivable. And Méndez would not be a moral agent in the traditional sense of the term ... The difference between Marlow and Myshkin does not consist solely of [whether they are conscious of their reasons for acting]. This means that the consciousness of one's reasons for acting is not the key to the

> distinction between moral subjecthood and moral agency. (Horta 2012, 93-4, my translation)[37]

If we understand Myshkin (M3) as we set out in the introduction to this section, we can see that he is quite close to (if not the same as) Horta's Méndez. The fact that Myshkin (M3) is unconscious of his reasons for acting means, as we saw, that he cannot access the operations that have led him to have a certain motivation and, as a result, cannot critically scrutinise it. But he *is* conscious of his motivations insofar as being motivated to act by an emotion implies being phenomenally conscious of that emotion. So, as Horta says of Méndez, he is conscious of his motivations but is unable to critically scrutinise them. If we understand "consciousness of one's reasons for acting" as the ability to access the operations of one's moral module and scrutinise its deliverances, then we can see how this is indeed the key to the distinction between moral subjecthood and moral agency.

A further objection that Horta presents has to do with Rowlands' realist position towards morality. Rowlands requires of a moral subject that it should have a sensitivity to at least some of the good- or bad-making features of situations, and he therefore presupposes that there are, in fact, certain features of situations that make them objectively good or bad. Horta

[37] "Rowlands concluye que lo único en lo que Marlow se distinguiría de Mishkin sería en el punto (iv). Marlow sí sería consciente de sus razones para actuar.

Hay motivos para pensar, sin embargo, que la diferencia entre Mishkin y Marlow es mayor que esa. Consideremos el caso de alguien que también tiene subjetividad moral, Méndez. Supongamos que Méndez actúa igual que Mishkin y Marlow, y que al igual que Marlow es consciente de sus razones para actuar. Supongamos, sin embargo, que Méndez es incapaz de llevar a cabo evaluación crítica alguna de sus motivaciones para actuar. Es consciente de ellas, pero se deja llevar totalmente por ellas sin poder llevar a cabo ninguna reflexión al respecto.

Pues bien, creo que Méndez es concebible. Y Méndez no sería una agente moral en el sentido en el que tradicionalmente se ha usado el término ... . La diferencia entre Marlow y Mishkin no consiste solamente en [la consciencia de las razones para actuar], pues. De manera que la consciencia de las razones para actuar no es la clave para la distinción entre subjetividad y agencia moral."

argues that this "is problematic simply because moral realism is false" (Horta 2012, 95, my translation).[38] With respect to this, we could say, firstly, that Horta's categorical assertion that moral realism is false is perhaps too hasty, if we take into account that this is still an ongoing debate in meta-ethics. Secondly, while Rowlands' account is indeed dependent upon some form of moral realism, it does not require the truth of one of the stronger versions of moral realism. Less stringent meta-ethical positions, such as Simon Blackburn's quasi-realism (Blackburn 1993), could also provide the adequate conditions for Rowlands' account of moral subjecthood.

To conclude this section, let us have a look at Mikel Torres Aldave's commentary (Torres Aldave 2012). He argues that the concept of moral subject is a "useless" and "dangerous" one. To make his point, he presents three objections. In his first objection, he argues that possessing moral emotions is not enough to be a moral being. To defend this claim, he offers the following argument:

> I have no problem assuming that animals are not only conscious beings, but also beings who are conscious of their intentions, motivations and emotions, as well as beings capable of thinking about the most adequate way of achieving their purposes. But one thing is for animals to have these capacities, and a very different thing is for them to be able to choose their intentions, motivations and emotions ... Besides asking ourselves how to obtain what we desire or that towards which our emotions move us, we humans have the capacity to ask ourselves if the mere act of desiring it or experimenting certain emotions is a good enough reason to choose that course of action among other possible ones ... The capacity to evaluate one's intentions and emotions turns out, from this point of view, to be a key element in guaranteeing the

[38] "[P]odría concluirse que el enfoque que hace Rowlands de la cuestión es problemático simplemente porque el realismo moral es falso."

> moral character and, therefore, the normativity of motivations and actions. (Torres Aldave 2012, 110-1, my translation)[39]

With this argument, Torres Aldave is merely reiterating the Kantian/Korsgaardian argument that Rowlands already replied to in his paper. Even Torres Aldave acknowledges this:

> I suppose Rowlands' answer to these possible objections would probably consist in pointing out that I've done nothing more than reiterate, although in a much less subtle and refined way, the arguments he criticises in his article. (Torres Aldave 2012, 114, my translation)[40]

Rowlands responded to this sort of objection by pointing out that this "capacity to evaluate one's intentions and emotions," which Torres Aldave correctly attributes to humans, in no way guarantees that we have control over the course of action that we end up following. This is so because the process of moving from the first-order level of our motivations to the higher-order level of critical scrutiny does not ensure control. The issue of whether we have control or not which arises at the first-order level will, as we saw, be merely reiterated at higher-order levels, unless we attribute some sort of miraculous power to the process of critical scrutiny.

[39] "No tengo problemas en asumir que los animales son no solamente seres conscientes, sino seres conscientes de sus intenciones, motivaciones y emociones, así como seres capaces de pensar acerca de la forma adecuada en que conseguir sus propósitos. Pero una cosa es que los animales tengan estas capacidades y otra bien distinta que sean capaces de elegir libremente sus intenciones, motivaciones y emociones. ... Aparte de interrogarnos acerca de cómo conseguir lo que deseamos o aquello a lo que nuestras emociones nos impulsan, los humanos tenemos la capacidad de preguntarnos si el simple hecho de desearlo o experimentar ciertas emociones es una razón lo suficientemente buena como para elegir un curso de acción entre otros posibles ... La capacidad de evaluar las intenciones y las emociones resulta, desde este punto de vista, un elemento esencial para otorgar carácter moral y, por tanto, normatividad a los motivos y las acciones."

[40] "Supongo que la respuesta de Rowlands ante estas posibles objeciones consistiría, probablemente, en señalar que no he hecho nada más que reproducir nuevamente, aunque de forma mucho menos sutil y refinada, los argumentos que él critica en su artículo."

Granting that Rowlands might not be convinced by this argument, Torres Aldave moves on to his second objection:

> the arguments against the control over moral motivation and actions on behalf of humans, presented in [Rowlands' paper] with the aim of solving the first problem [*sc*. the Kantian idea that without scrutiny there is no control, and thus no normativity and no morality], seem to undermine, in an unsolvable way, the idea of moral agency. (Torres Aldave 2012, 106, my translation)[41]

Torres Aldave is worried that Rowlands' framework cannot accommodate the concepts of freedom and responsibility, and thus the idea of moral subjecthood would be a "dangerous" one because, in its defence, Rowlands has wound up eliminating the possibility that we might be moral agents. I have already discussed how Rowlands deals with this problem in his book, namely by arguing that our ability to metacognise grants us an understanding of our actions that allows us to be held responsible for them. To avoid further repetition, I will now make a different point.

Rowlands makes it very clear in his book that his argument against control only affects the notion of control gained through scrutiny, which is the form of control that has traditionally been denied to animals (Rowlands 2012b, 235). This suggests that there is no reason to completely abandon the idea of control. The drunken driver is less in control of his actions than the sober one, not because he can't scrutinise his motivations, but because of the effects that alcohol has on his system. We are more in control of our actions when we are not drunk, and

[41] "[L]os argumentos en contra del control de la motivación y las acciones morales por parte humana, presentados en el texto con el objetivo de superar el primer problema, parecen socavar, de forma irreparable, la idea de agencia moral."

also when we are not being threatened or coerced. We also have an ability to inhibit certain impulses, we don't follow fixed action patterns, we can learn, and we adapt our behaviour to fulfil our desires. These forms of control (let's refer to them as control*, to distinguish them from higher-order forms of control gained through scrutiny) cannot, however, be plausibly denied to animals, at least not to the ones who are possible candidates for moral subjecthood.

With this in mind, I suggest that the best way to think of moral agency is as a spectrum, where both the amount of understanding and the amount of control* one has over one's motivations plays a role in determining at which point of the spectrum one is standing at one particular point in time. Animals who are moral subjects may have a small degree of moral agency whenever they are not, for instance, performing their actions under the effect of drugs or coercion. Their lack of understanding, however, prevents them from attaining the higher degrees of moral agency to which we attach full-blown moral responsibility. This is coherent with our attitudes towards animals. We are, for instance, angry at our dog when he chews on our slippers, but not when he vomits on the rug. The reason for not being angry in the second case is that we know he couldn't help it, he wasn't in control* of his action. In the first case, we consider that being angry is warranted to a certain extent, for the dog was in control* of his action, but we can't *really* blame him because we know that, though he may have *chosen* to chew on our slippers, he doesn't really *understand* the wrongness of what he just did. This is also coherent with what Rowlands argues in his book:

> If we think of agency as a spectrum whose left-hand side corresponds to the low, and runs to high on the right, then animals might be thought of as occupying a position on the far left, with normal, adult humans clustering toward the right. It is all a matter of degree, and in the case of animals, the degrees involved are small enough that, if we were thinking in all-or-nothing terms, we would be inclined to say they are not agents. (Rowlands 2012b, 241)

The last objection presented by Torres Aldave is similar to the one introduced by Longueira Monelos. That is, he considers that there is nothing to gain, from a practical point of view, by introducing the notion of moral subjecthood. This is a rather important objection, and is also one mentioned by Ángel González-Torre (2012, 139) and Kepa Tamames (2012), so it is worth considering again. I already pointed out one practical implication that the notion of moral subjecthood can have, namely, that it may raise awareness of the implausibility of supposing a discontinuity between the nature of humans and animals, and thus undermine any claims of human superiority that may be used to justify the mistreatment of animals. Some further practical implications are, inadvertently, suggested by Torres Aldave himself:

> The distinction between moral agents and patients is a fruitful one, both in theory and in practice. It serves, at an ontological level for example, to characterise and explain the different nature of different types of beings, as well as to classify them. At an ethical-political level, the distinction is fundamental when it comes to determining the moral responsibility of individuals and, for instance, the penal consequences of their actions. Also, the distinction moral agents/patients serves as well to tackle the type of relationship that obtains, and should obtain, between them. (Torres Aldave 2012, 117-8, my translation)[42]

By substituting the term 'moral patient' with 'moral subject' in this paragraph, we can easily devise other practical implications of Rowlands' notion. Firstly, this new concept can

[42] "La distinción entre agentes y pacientes morales es, tanto en la teoría como en la práctica, fructífera. Sirve, a nivel ontológico por ejemplo, para caracterizar y explicar la distinta naturaleza de distintos tipos de seres, así como para realizar clasificaciones entre los mismos. A nivel ético-político, la distinción resulta fundamental a la hora de determinar la responsabilidad moral de los individuos y, por ejemplo, las consecuencias penales de sus acciones. Además, la distinción agentes/pacientes morales sirve también para abordar el tipo de relaciones que mantienen y deben mantener unos con otros."

undoubtedly serve to characterise and classify different beings. Secondly, at times it may also be fundamental to determine whether a being is a moral subject or a moral agent in order to decide what her moral responsibility is and what penal consequences her actions should have. The notion of moral subjecthood allows us, for example, to understand that a dog who bites a child has done something wrong, but does not grasp the wrongness of his action and so should arguably be trained to prevent future incidents, instead of killed. Lastly, the distinction also helps to establish what kind of relationship should exist between moral agents and moral subjects. For instance, a case could be made to argue that dog owners have a responsibility to ensure that the virtues of loyalty, patience, and compassion, towards which most dogs are naturally inclined, get a chance to flourish in them. Ensuring an animal gets treated with dignity and respect is more than simply preventing her from feeling pain and suffering, contrary to what Tamames suggests (Tamames 2012, 144). The better we know the nature of a certain animal, the better we can tailor our relationship with her to fit her qualities and allow them to flourish, a point I shall return to in chapter 6. For now, we can conclude that the reason why the concept of moral subjecthood may initially appear useless, is probably and quite simply because we are not used to having it as part of our conceptual repertoire.

2.4. SUMMARY

The aim of this chapter has been, on the one hand, to give an overview of the debate on animal morality, and, on the other hand, to defend the notion of moral subjecthood as a means for advancing this debate. We have seen, firstly, that there is a considerable amount of evidence that suggests that some animals may sometimes behave on the basis of altruistic and cooperative tendencies, and may even possess a sense of what is fair and what is not. This evidence takes three forms: experimental, observational, and anecdotal. All together it builds a preliminary case for the idea that some animals may possess moral motivations. However, we have seen that philosophers and scientists are often reluctant to seriously consider this possibility. This reluctance is commonly defended by an appeal to an argument that is problematic, both because it assumes a highly intellectualistic conception of morality, and because it depends on the outcomes of ongoing empirical investigations. Furthermore, we have also seen that, at the basis of these authors' denial of animal morality, lies the conflation of three separate qualities, all of which are relevant to morality: the possession of moral motivations, the capacity to make moral judgements, and the possibility of being held morally responsible for one's actions. I have argued that the conflation of these three ideas amounts to a fallacy—what I have termed the *amalgamation fallacy*. The best way to sidestep this fallacy is by incorporating the notion of *moral subjecthood*, introduced by Mark Rowlands, into the debate. I have shown that there are powerful reasons that support the coherence of this notion, and that the objections that have been raised against it can be satisfactorily addressed. The question of animal morality should, therefore, no longer be whether animals can be full-blown moral agents in the same way a normal adult human is. Instead, we should be asking whether animals can be moral subjects.

3. MINDREADING ANIMALS?

In the previous chapter, I gave an overview of the debate on animal morality, a debate that is alive and showing no signs of resolution in the near future. Despite the growing empirical evidence, we saw that several philosophers and scientists refuse to seriously consider the possibility that animals may possess morality. I have argued that this refusal commonly occurs because authors have fallen prey to the amalgamation fallacy. This means that morality is mistakenly understood by these authors as a whole with three parts that cannot be disentangled: (1) the capacity to behave on the basis of moral motivations, (2) the ability to make moral judgements, and (3) the property of having moral responsibility over one's actions. I have argued that, in order to advance the debate, it is necessary that we sidestep the amalgamation fallacy by adopting a moral category that pertains only to the capacity to behave morally—the category of *moral subjecthood*. Mark Rowlands, as we saw, was the first to introduce this notion (Rowlands 2011; 2012a; 2012b). Traditionally, (mere) moral subjecthood has been tacitly assumed as a logical impossibility, for it is a category that implies the existence of moral motivations without moral responsibility. Rowlands' defence of the coherence of this notion turns on a demonstration that the capacity to engage in metacognition (which is a key ability for both moral judgement and moral responsibility) is not necessary for one's motivations to be moral in character. The thesis that morality can occur in the absence of metacognition is at the very core of Rowlands' position.

Metacognition is a form of higher-order thought, for it enables one to entertain thoughts about one's own mental states. The other great form of higher-order thought is what is usually termed ‘mindreading’ or ‘theory of mind,’ which consists in the conceptualisation and attribution of mental states to others. If one has mindreading capacities, one can entertain thoughts about the mental states of others. My purpose throughout this dissertation is to develop Rowlands’ position by defending the thesis that morality does not require the presence of mindreading abilities, either. That is, I’m going to argue that a being does not have to be capable of attributing mental states to others in order to qualify as a moral subject. As we shall see in this chapter, and later in chapter 5, this is a claim that will go against several well-established intuitions.

The previous chapter was intended as a first approximation to the topic of animal morality. This chapter, in turn, contains a first approximation to the question of animal mindreading. It is divided into three parts. In the first part, I sketch the reasons behind the common assumption that morality requires mindreading capacities. In the second part, I give a historical overview of the emergence and development of the mindreading debate. In the last part, I introduce the behaviour reading hypothesis—a challenge to the animal mindreading debate that is most commonly associated with Daniel Povinelli and will have an important role to play in later chapters.

3.1. MINDREADING AND MORALITY

Moral philosophers and moral psychologists tend to assume that being moral requires possessing mindreading capacities. This is an assumption that is rarely made explicit, but that, as we shall see, underlies a good part of the debates on moral psychology, as well as the most important theories in normative ethics. The reason why it is rarely made explicit probably stems from its largely intuitive character. Indeed, the claim that one cannot be moral unless one possesses the capacity to grasp (1) that others can feel pain and suffer, and (2) the impact that one's actions can have on other beings' mental lives, seems to be obviously correct. Despite its intuitive character, this is an assumption that I intend to question, and will indeed do so in chapter 5. At this point, however, it's important to explore the reasons behind the pervasiveness of this assumption. There are largely two ways in which mindreading is thought relevant for morality. On the one hand, it is thought relevant for the possession and adequate exercise of moral judgement and, on the other hand, it is understood as necessary for one's motivations and actions to be moral in character. Let us now explore each of these in turn.

3.1.1. MINDREADING AND MORAL JUDGEMENT

Moral judgement, understood as the process whereby we scrutinise our own and others' motivations and behaviour in order to determine their moral status, is commonly thought to engage our mindreading capacities. This is a claim that is backed up by a good deal of empirical research. Several studies from developmental psychology suggest that a higher development of children's mindreading capacities is positively correlated with more sophisticated performance in moral reasoning and moral judgement tasks (Baird and Astington 2004; Dunn, Cutting, and Demetriou 2000; Piazzarollo Loureiro and de Hollanda Souza 2013; Lane et al. 2010; Cushman et al. 2013). Moral psychologists have found evidence in adult humans of a strong reliance of moral judgements upon the attribution of mental states, especially intentions, to others (Cushman 2008; Darley, Klosson, and Zanna 1978; Woolfolk, Doris, and Darley 2006). Studies in social neuroscience have also found that brain areas that are associated with mindreading are often activated during moral judgement and moral reasoning tasks (Young et al. 2007; Young and Saxe 2008; Harenski et al. 2009; Bzdok et al. 2012). With this sort of research in mind, several authors have postulated that mindreading capacities play a crucial role in (human) moral judgement.

Steve Guglielmo, Andrew Monroe and Bertram Malle (2009), for instance, have defended the following:

> When people make a moral judgment they evaluate an agent's behavior in light of a system of norms. Such evaluations of behavior are enmeshed with inferences about what was in the agent's mind before, while, and even after performing the behavior. If folk psychology is the system of concepts and processes that enable a human social perceiver to make such inferences from behavior, then folk psychology lies at the heart of moral judgment. (Guglielmo, Monroe, and Malle 2009, 449).

Guglielmo *et al.* are using the term 'folk psychology' to refer to "the system of concepts and processes" (i.e. the cognitive mechanisms) that enable "inferences about what was in the agent's mind before, while, and even after performing the behavior" (i.e. the attribution of mental states). The thesis that the authors are putting forward thus amounts to the claim that mindreading capacities lie "at the heart of moral judgment." The reasoning behind their claim is that mindreading is involved whenever individuals engage in moral judgements such as "blaming, assigning responsibility, feeling resentment or sympathy" (Guglielmo, Monroe, and Malle 2009, 449). Thus, whenever people blame someone for something she has done, hold her responsible for it, or feel resentment or sympathy at her, there is a reasoning behind this process that involves the attribution of mental states. That is, people generally have to arrive at the conclusion that the target has undergone certain requisite mental states (beliefs, desires, intentions, knowledge, and so on), for them to feel that blame, responsibility, resentment, sympathy, and so forth, are warranted. For instance, in the case of the moral judgements involved in blame, people require the agent in question to have *intentionally* engaged in harmful, or otherwise wrong, behaviour. In order to judge that a certain behaviour was indeed intentional, and thus blameworthy, people require, amongst other things, "evidence for the agent's *desire* for an outcome, *beliefs* about the action in question leading to the outcome, the *intention* to perform the action, *awareness* of the act while performing it" (Guglielmo, Monroe, and Malle 2009, 451, their emphasis), the evaluation of which engages our mindreading capacities.

Kurt Gray, Liane Young and Adam Waytz (2012) have asserted that "the essence of moral judgment is the perception of two complementary minds—a dyad of an intentional *moral agent* and a suffering *moral patient*" (Gray, Young, and Waytz 2012, 101, their emphasis). The authors contend that what all moral judgements have in common is the involvement of what they call 'mind perception.' They consider this to be different from, and more basic than, an exact understanding of another's mental contents (Gray, Young, and Waytz 2012, 103). Nevertheless, insofar as mind perception involves the attribution of mental states to others (even if in a fuzzy and indeterminate manner), we can consider it a form of mindreading, in the broad sense in which I've chosen to understand this term. The essence of morality, they argue, is "the combination of harmful intent and painful experience" (Gray, Young, and Waytz 2012, 106), and so, at the heart of most moral judgements, lies the perception of a dyad of mental states: harmful intent on behalf of the agent, and painful experience on behalf of the patient. If one lacks the capacity to mindread (or, as they call it, 'perceive minds'), then one cannot engage in proper moral judgements. The authors illustrate this with the example of two disorders that affect moral judgement in humans. On the one hand, autistic patients have a difficulty in perceiving agency (a form of mental state attribution), and correspondingly fail to properly understand moral agents. On the other hand, psychopathic patients have a difficulty in perceiving experience or sentience (also a form of mental state attribution), and thus fail to properly understand moral patients (Gray, Young, and Waytz 2012, 105).

In a later paper, Liane Young and Adam Waytz (2013) have asserted that "[m]orality depends critically on our ability to attribute minds to entities that engage in moral actions (towards ourselves and others) and the entities that experience these actions (our own actions and others')" (Young and Waytz 2013, 93). In particular, they claim that the main role of our mindreading capacities can be found in moral cognition, as they constitute a fundamental mechanism when it comes to understanding the actions of moral agents, predicting those actions of others that will affect us, as well as determining who is our ally and who our enemy (Young and Waytz 2013, 93). After reviewing research done by moral psychologists, the authors assert that when we assign blame to a certain agent, the mental states she has undergone appear to be much more important than the outcomes of her action, and even more important than whether the agent could have done otherwise or not, so that "[m]ental state factors are clearly at the forefront of our minds when we're making moral judgements" (Young and Waytz 2013, 95).

In a paper discussing the requirements for the creation of moral machines, Paul Bello and Selmer Bringsjord (2013) have argued that "mindreading plays an integral role in [human] moral cognition," for "[w]ithout the ability to consider the beliefs, desires, intentions, obligations and other mental states of our confederates, much of the richness of human moral cognition evaporates" (Bello and Bringsjord 2013, 256). If we want to create moral machines that will interact adequately with human beings in all their complexity, these authors argue, we must find a way of implementing mindreading capacities in these artificial moral beings.

These examples serve as an overview of the relation between moral judgement and mindreading, as it is viewed from the perspective of disciplines such as experimental philosophy, moral psychology, and social neuroscience. The common claim is that *people's moral judgements involve mindreading*. My interest, however, is in arguing that *mindreading capacities are not a necessary requirement for morality*. When these authors claim that "at the heart of morality lies folk psychology" (Guglielmo, Monroe, and Malle 2009, 449), that "mind perception is the essence of morality" (Gray, Young, and Waytz 2012, 101), or that "mind attribution is for morality" (Young and Waytz 2013, 93), they are not directly challenging the thesis I want to defend. This is, firstly, because their claims are ones that refer to the issue of moral judgement, but not to the issue of moral behaviour more generally. What they are asserting is that, at the heart of our capacity to engage in moral judgement, lies our capacity to attribute mental states to others. They generally do not address the question of whether behaving in a moral manner, or on the basis of motivations that are moral in character, requires the presence or involvement of mindreading capacities—which is precisely where my interest lies.

The second big difference between the claim Guglielmo et al. (2009), Gray et al. (2012), and Young & Waytz (2013) defend and the one I want to dispute is that these authors are stating a *contingent* fact about (human) morality, and not describing a *necessary* one. Their arguments are made after empirical examinations of people's moral judgements in different circumstances. What is being argued is not that mindreading *must* be involved in order for one to perform a moral judgement, but rather, that mindreading is *actually* involved whenever people judge the morality of others' behaviour. The idea that (normal adult) humans make use of their mindreading capacities whenever they engage in moral judgements is not one that I intend to question. Even if these authors were defending the claim that humans engage in the attribution of mental states to others whenever they *behave* morally (as seems to be suggested by Young & Waytz, 2013, pp. 97—100), I would not be interested in disputing that claim, either. This is because, if it were true that all human moral behaviour involves mindreading *de facto*, it would not follow that moral behaviour *necessarily* involves mindreading. My interest lies precisely in arguing that mindreading does not have to be involved in order for a behaviour or a motivation to count as moral, and moreover, that an individual does not have to possess mindreading capacities to be a moral subject. These scholars do not take a stand on this issue. However, as we shall see in the following section, the same cannot be said of many moral philosophers.

3.1.2. MINDREADING AND MORAL BEHAVIOUR

This section will explore the reasons behind the common assumption that moral behaviour requires mindreading capacities. This is an idea that is presupposed by most of the moral philosophers that adhere to any of the three main theories in normative ethics: consequentialism, deontology, and virtue ethics. These theories are concerned with two different endeavours. Firstly, they provide an account of the nature of the Good or the Right, and secondly, they provide instructions on the morally correct way to act in any given situation. The claim that morality requires mindreading applies to this second aspect. That is, from the viewpoint of most consequentialists, deontologists, and virtue ethicists, the requirements to act in the morally correct way *cannot be met* without an ability to grasp the mental states of others. This is an assumption that is presupposed from an *a priori* standpoint. The idea is that there is something about morally correct behaviour that means it *necessarily* requires the presence of mindreading capacities. It is not, therefore, a claim about the contingent character of moral behaviour. The thesis is not that moral behaviour *de facto* involves mindreading, but that *behaviour cannot be moral unless the individual who engages in it possesses mindreading capacities*. Let us have a closer look at how this claim is present in these three ethical theories.

For consequentialists, "choices—acts and/or intentions—are to be morally assessed solely by the states of affairs they bring about" and "whatever choices increase the Good, that is, bring about more of it, are the choices that it is morally right to make and to execute" (Alexander and Moore 2012). Thus, from this point of view, being capable of acting in the morally correct way requires, firstly, the capacity to reflect upon one's motivations in order to choose the one that will increase the amount of Good in the world. This means that consequentialists are going to require that moral beings possess metacognitive capacities, insofar as one's motivations are mental states and reflecting upon them requires metacognition. What is important for our purposes is, however, the fact that the Good is, more often than not, defined in terms of others' mental states, and so choosing whichever one of our motivations will increase the Good not only requires metacognitive capacities, but also mindreading abilities. This is most obvious in the case of utilitarians, who understand the Good in terms of "pleasure, happiness, desire satisfaction, or 'welfare' in some other sense" (Alexander and Moore 2012). Other consequentialists may emphasise other forms of Good apart from positive mental states, but generally, from a consequentialist perspective, being moral requires having the capacity to assess the outcomes of one's actions, and this will always require having an understanding of the impact that they will have on other beings' mental lives. The idea that moral behaviour requires mindreading is, therefore, at the very core of (at least most forms of) consequentialism.[1]

Deontologists de-emphasise the importance of consequences and focus on the acts themselves insofar as they exemplify a given motivation or moral principle. Rather than stressing the importance of increasing the amount of Good in the world, deontological theorists stress the importance of performing the right actions for the right reasons. The requirement that one's actions be performed *for the right reasons* leads deontologists, as well, to emphasise the importance of an assessment of the circumstances in which our actions take place. On the one hand, a lot of emphasis is placed on the importance of critically scrutinising one's motivations, to ensure that one chooses those that are in accordance with the appropriate duties and responsibilities, and so, metacognition is also required here. But at the same time, deontologists, especially those close to the Kantian tradition, stress the importance of the intentions behind our actions and the need for them to be in accordance with an understanding of others as ends-in-themselves and not merely as a means to our purposes. The claim that moral behaviour requires mindreading is implicit here insofar as a creature unable to understand others as possessors of mental states could never do the right thing *for the right reasons*. Perhaps such a creature could behave in apparent accordance with a certain duty, but she couldn't have chosen such behaviour for the right reasons (i.e. with a mindful understanding of the rights of others and their nature as ends-in-themselves), so from a deontological perspective we could never consider that she behaved in the *morally correct* way.

Lastly, virtue ethicists focus on the importance of having the right character traits, rather

[1] Some consequentialist authors adopt an externalist perspective in order to argue that being moral requires behaving on the basis of motivations or character traits that systematically produce good consequences, independently of the agent's ability to assess the consequences themselves (Driver 2006; Rowlands 2012b). This is a form of consequentialism that does not require metacognition or mindreading (see sections 2.3.2 and 5.3.5).

than performing the right actions or producing the best consequences. Nevertheless, possessing a certain virtue is far from an easy task. It consists of being “a certain sort of person with a certain complex mindset,” and requires “the wholehearted acceptance of a certain range of considerations as reasons for action” (Hursthouse 2013). Being an honest person, for instance, requires telling the truth in a consistent manner, but it is far from just that. If an individual always told the truth out of an indifference towards other people’s feelings, she would not be virtuous. Any virtue requires persistently performing the right actions for the right reasons, in the right manner, to the right people, at the right time, etc. A careful and consistent assessment of one’s motivations and the circumstances that surround our actions is needed to acquire a virtue. Without an ability to grasp the mental states of those around us, this requirement cannot be met.

Despite its pervasive character, the thesis that moral behaviour requires mindreading is seldom made explicit.[2] Nevertheless, one can occasionally come across a reference to it. The reader may recall Marc Hauser’s (2001) six arguments against animal morality, which we saw in the previous chapter. Mindreading capacities were amongst the requirements that Hauser established for truly moral behaviour. In his own words:

> [I]f animals are moral agents, they must understand how their actions will impact on the feelings and thoughts of other individuals of their species, and take these consequences into consideration when planning an action. If animals lack the capacity to attribute mental states to others, then they are creatures of the present. This is not to say that they lack the capacity to anticipate or think about the future …

[2] An exception is the debates surrounding the topic of empathy, where the link between mindreading and moral motivation is often discussed. I will ignore this for now, as the following chapter is devoted entirely to the relationship between mindreading and empathy.

> What they lack is the capacity to think about how their own actions might change the beliefs of others or cause them pain or happiness, and how risky behaviors such as an attack on a new mother or a competitor might result in their own death, the termination of life and all that it brings. (Hauser 2001, 313)

Hauser's claim is that, without the capacity to attribute mental states to others, one cannot understand the impact that one's actions can have on other beings' mental lives, and thus cannot make properly informed decisions on how to act. The absence of mindreading capacities, Hauser argues, restricts an individual to the present, for it prevents a proper evaluation of the consequences of her actions. Gregory Peterson (2000) follows a similar train of thought when he suggests that mindreading may be necessary for morality. Mindreading capacities may be crucial for morality, he argues, because of their relevance in our ability to "rationally deliberate actions and their consequences" (Peterson 2000, 475). Peterson and Hauser thus seem to be explicitly stating the reason behind the mindreading requirement that we saw represented in the three great ethical theories: without an ability to attribute mental states to others, we cannot engage in a rational scrutiny of our motivations, our behaviour, and/or their consequences—a scrutiny that the three ethical theories, for different reasons, consider necessary for morality.

In this argument, mindreading is portrayed as essential for morality for the same reason that metacognition is commonly deemed necessary as well. Both abilities are understood as a tool to engage in a proper scrutiny of our motivations that will ensure that we act in accordance with what is morally right. Rowlands' critique of the role of scrutiny in moral motivation (which we saw extensively in section 2.3.2) can be easily extended to this understanding of the importance of mindreading for morality. The account of moral behaviour he constructs, as we saw, is one that is entirely independent of a capacity to subject one's motivations to critical scrutiny. Mindreading as a tool for critical scrutiny thus has no role to occupy in his framework. It would be unwise, however, to think that mindreading can be set so easily aside, for there are powerful reasons for thinking that mindreading, specifically and independently of its role in a critical scrutiny of one's motivations, is necessary for our behaviour to qualify as moral. In particular, in the literature we can find at least four independent arguments to this effect, offered by Cheryl Abbate (2014), Florian Cova (2013), Shaun Nichols (2001), and Jelle de Boer (2011). Let us now have a look at each of them.

Cheryl Abbate (2014) has defended the idea that moral agency requires "the capacity to form beliefs or thoughts about the mental states of others" (Abbate 2014, 10). The reasoning behind this claim is that, without an ability to attribute mental states to others, "one cannot be held responsible for affecting another's mental states since one is unaware *that* one's actions affect others" (Abbate 2014, 10, her emphasis). An awareness of how our actions affect others' mental states is thus a necessary condition for being morally responsible for our actions. The capacity to mindread is, in turn, a necessary (though not sufficient) condition for understanding "the significance (the rightness or wrongness) of affecting another's mental state," which is also deemed necessary for moral responsibility (Abbate 2014, 13—4). This claim can also be found in Hauser (2001, 312) and Francescotti (2007, 246).

Florian Cova (2013) has argued that "[a]ll it takes to be a moral agent is ... to be able to act according to what we attach importance to and a bit of theory of mind" (Cova 2013, 128). Cova establishes that what is required for a minimal conception of moral agency is the presence of *care* for others, where caring for someone, at a minimum, entails being "able to understand that this person has *interests* — that some things have (positive or negative) values for this person," as well as assigning "an intrinsic (i.e. non-instrumental) positive value to the fact that this person's interests are preserved and augmented, and [giving] an intrinsic negative value to the fact that this person's interests are damaged" (Cova 2013, 127). Moral behaviour cannot occur without mindreading capacities because it "requires understanding that others have interests and the capacity to be motivated by this understanding (that is: to act in accordance with how much I care about others)" (Cova 2013, 128).

Shaun Nichols (2001) has argued that "a person cannot have the capacity for altruistic motivation without some capacity to attribute negative affective or hedonic states to another"

(Nichols 2001, 436). Keeping in line with his commitment to experimental philosophy (see Knobe and Nichols 2008), Nichols makes an empirically-informed argument to support his claim. He considers the most basic form of altruistic behaviour: helping or comforting behaviour directed at another individual in distress. In the absence of mindreading capacities, Nichols argues, an individual in a situation like that would "only help when it's easier to help than it is to escape," and yet, empirical research shows that "in core cases of altruism, people often prefer to help even when it's easy to escape" (Nichols 2001, 431; see also Batson 1991). Altruistic behaviour needs to be motivated by "an enduring *internal* cause" to ensure that "escape is not an adequate alternative" to helping. This sort of enduring internal cause can only be provided by a representation of the other's negative mental state, for "[i]f altruistic motivation is triggered by a representation that the target is in pain," the helping reaction is guaranteed, as "merely escaping the perceptual cues of pain won't eliminate the consequences of the enduring representation that another is in pain" (Nichols 2001, 436).

Lastly, Jelle de Boer (2011) considers that a "moral society" requires "mutual beliefs," so that those who behave morally do so, not only because they believe that it is what they *should* do, but also because they believe that everybody else in the society also believes that that is what one should do, in those circumstances. A moral society emerges when "[w]hat should be done is a moral fact," which means that "everybody believes this about each other." It would not be enough for it to be the case that "each individual simply believes what should be done, each on his own, with all merely converging on the moral truth," for this would not secure the appearance of a moral norm. The latter emerges, according to Boer, when there is a mutual consensus with respect to the norm, and this consensus is common knowledge (de Boer 2011, 900—903). Obviously, one cannot hold a belief about what others believe without mindreading capacities, and so this requirement is present here too.

These authors' reflections allow us to elaborate a set of four different conditions that, if found necessary for moral behaviour, would entail that it requires mindreading capacities, independently of the role of this ability in a critical scrutiny of our motivations:

1. *The responsibility condition:* Being moral requires understanding both *that* and *how* our actions affect others' mental states, as well as the moral significance of it. Without this sort of understanding, Abbate (2014) and Hauser (2001) argue, one cannot be held morally responsible for one's actions.
2. *The caring condition*: Being moral requires caring for others, where this means understanding that they have interests and assigning a positive value to the fulfilment of these interests. Without mindreading capacities, according to Cova (2013), there can be no care.
3. *The motivation condition*: The motivations behind moral behaviour have to be what Nichols (2001) terms "enduring internal causes," which means they must be triggered by a representation of another's negative mental state, in order to guarantee that they cause helping, and not fleeing behaviour.
4. *The reciprocity condition*: Mindreading capacities are necessary to ensure that moral norms are shared among the members of a society. This condition has been formulated by de Boer (2011), who argues that a moral norm exists only when individuals, not only behave in accordance with it, but also have a belief to the effect that everyone else believes the norm should obtain.

These four conditions are not meant to be understood as a set. Rather, each of them separately constitutes a possibly necessary condition that must be met by an individual's behaviour for it to count as moral. So, they are not *collectively* necessary, but rather, if any of them were found to be truly a necessary condition for moral behaviour, this would entail the automatic exclusion of non-mindreaders from the realm of moral subjects, for it would mean that behaving morally *necessarily requires* mindreading capacities. In chapter 5, I will have a closer look at each of these conditions, and the arguments that support them. My aim will be to show that none of these authors succeed in establishing a necessary connection between mindreading and moral behaviour. This will allow me to conclude that mindreading capacities are not required for an individual to be a moral subject. But first, it is important to convey the reason why these four conditions pose a potential threat to the case for animal morality. Accordingly, the rest of this chapter will be devoted to illustrating the current situation in debates on animal mindreading.

3.2. A BRIEF HISTORY OF THE ANIMAL MINDREADING DEBATE

This dissertation's main claim is that mindreading capacities are not necessary for morality. In the previous section, we have seen that this thesis challenges an assumption that is pervasive throughout normative ethics, as well as backed up by several independent theoretical considerations. There are powerful reasons for thinking that morality requires mindreading. This is problematic for the case for animal morality, insofar as the fruitfulness that we saw in the animal morality research is somewhat lacking in investigations on animal mindreading. As we shall see in this section and the following one, there are very few experiments that have yielded positive results with respect to animal mindreading, and many scholars doubt the validity of what little evidence we have. Arguing for the logical independence of morality and mindreading will be crucial, if we want to prevent the discouraging results in the animal mindreading debate from trumping the possibility of concluding that some animals are nevertheless moral. To understand this more carefully, it is necessary to begin by having a look at the history of the mindreading debate.

The mindreading debate began in 1978 with a paper published by David Premack and Guy Woodruff, entitled 'Does the chimpanzee have a theory of mind?' The term 'theory of mind' was coined in this paper and originally defined in the following way:

> In saying that an individual has a theory of mind, we mean that the individual imputes mental states to himself and to others (either to conspecifics or to other species as well). A system of inferences of this kind is properly viewed as a theory, first, because such states are not directly observable, and second, because the system can be used to make predictions, specifically about the behavior of other organisms. (Premack and Woodruff 1978, 515)

This paper triggered three separate debates that are still alive today:

1. The first of these debates concerns the question of whether any other animal, besides humans, possesses a so-called theory of mind. As we shall see, the answer to this question is still far from clear.
2. The second debate addresses the question of how this capacity develops in humans, at what age it first appears in us, and what sort of cognitive and emotional deficits come with an inappropriate development of this capacity. For instance, attempts have been made to explain disorders such as autism by appealing to an absence of theory of mind (e.g. Baron-Cohen, Leslie, and Frith 1985).
3. The third debate concerns the cognitive mechanisms that underlie our theory of mind. For some time, the choice stood between two different conceptions of these mechanisms. On one side, stood the so-called *theory-theory*, which was inspired by Premack and Woodruff (1978)'s original definition of theory of mind, and "holds that when we mindread we access and utilize a theory of human behavior represented in our brains" (Ravenscroft 2010). On the other side, one could find the *simulation theory of mind*, which was originally introduced by Robert Gordon (1986), and "holds that we represent the mental states and processes of others by mentally simulating them, or

> generating similar states and processes" (Gordon 2009). Recent times have seen the development of hybrid approaches that encompass elements from both these theories (e.g. Nichols and Stich 2003; Goldman 2006). Additionally, innovative and unorthodox approaches to social cognition are starting to gain strength (e.g. Hutto and Gallagher 2008; Hutto 2012; Andrews 2012b).

This section will provide an overview of the first of these three debates—the one that addresses the issue of whether any nonhuman species possesses a theory of mind. Because I will not take a stand on the debate regarding which cognitive mechanisms underlie this capacity, I will avoid the use of the term 'theory of mind,' to sidestep its theory-theory connotations. Instead, I will continue to opt for the term 'mindreading,' and use it, as I have been doing until now, to refer to the attribution of mental states (emotions, sensations, beliefs, desires, intentions, and so on) to others. Correspondingly, the phrase 'mindreading capacities' or 'mindreading abilities' will be used to refer to whichever cognitive mechanisms enable an individual to engage in mindreading. In the literature, the term 'mindreading' and 'metacognition' are often used as synonyms, as some authors believe that the same cognitive mechanisms lie at the basis of the attribution of mental states both to oneself and to others (e.g. Carruthers 2009). I will, however, restrict the use of the term 'mindreading' to the attribution of mental states to *others*, and continue using the term 'metacognition' to refer to the attribution of mental states to *oneself.* I shall take no stand on whether metacognition and mindreading are dependent upon the same cognitive mechanisms, as it is not an important issue for the arguments I will develop in upcoming chapters.

Let us begin, then, this brief history of the animal mindreading debate by having a look at David Premack and Guy Woodruff's original experiment, which was published in *Behavioral and Brain Sciences* with an open peer commentary. Premack and Woodruff presented a chimpanzee named Sarah with several videotape recordings of a human experimenter trying to solve a certain problem, like reaching for a banana, escaping from a cage, or attempting to turn a heater on. After each viewing, Sarah had to choose from a pair of photographs that presented objects or actions, only one of which constituted a solution to the problem that the experimenter was facing in the video. She had virtually no problem in solving the puzzles she was presented with. The authors suggested the following as the best explanation for Sarah's outstanding performance in these tests:

> [T]he chimpanzee solves problems such as [these] by *imputing states of mind* to the human actor. In looking at the videotape, he imputes at least two states of mind to the human actor, namely, *intention or purpose* on the one hand, and *knowledge or belief* on the other. The chimpanzee makes sense of what he sees by assuming that the human actor wants the banana and is struggling to reach it. He further assumes that the actor knows how to attain the banana (Premack and Woodruff 1978, 518, my emphasis)

These experimental results were, therefore, presented as preliminary evidence for the presence of mindreading capacities in a member of a nonhuman species. However, the majority of authors who participated in the open peer commentary were very sceptical of this interpretation of the data. In order to pass the test, many argued, Sarah did not need to attribute mental states to the actor. She could have been solving the task by using "relatively simple associationistic and match-to-sample strategies" (Savage-Rumbaugh, Rumbaugh, and Boysen 1978, 556), or simply choosing "the photograph that represents what she would like to happen" (Burge 1978, 561).

Out of all the commentaries, the most influential one was perhaps Daniel Dennett's, who argued that Premack and Woodruff's experimental paradigm was unsatisfactory because it required extensive familiarity with the situation on behalf of Sarah, which meant that she could solve the task by making inferences from environmental circumstances, rather than through mental state attribution. A way to obtain good evidence of Sarah's mindreading capacities, Dennett suggested, would be to place her in a situation in which a certain reaction could only be expected of her if she were capable of representing the other individual's *false* belief. Dennett exemplified this with his famous *Punch and Judy* example:

> Very young children watching a Punch and Judy show squeal in anticipatory delight as Punch prepares to throw the box over the cliff. Why? Because *they know Punch thinks Judy is still in the box*. They know better; they saw Judy escape while Punch's back was turned. We take the children's excitement as overwhelmingly good evidence that they understand the situation—they understand that Punch is acting on a mistaken belief (although they are not sophisticated enough to put it that way). Would chimpanzees exhibit similar excitement if presented with a similar bit of play acting (in a drama that spoke directly to their "interests")? I do not know, and think it would be worth finding out for if they didn't react, the hypothesis that they impute beliefs and desires to others would be dealt a severe blow, even if all the [Premack & Woodruff] tests turn out positively, just because it can be made so obvious—obvious enough for four-year-old children—that Punch believes (falsely) that Judy is in the box. (Dennett 1978, 569, his emphasis)

We can know that a child is capable of attributing mental states to others, Dennett argues, when we see her laugh as Punch prepares to throw a box off a cliff. We know that she is capable of this attribution because the reason she finds the situation funny is that she *understands* that Punch *mistakenly believes* that Judy is in the box. Dennett suggested that we should test chimpanzees' mindreading capacities by placing them in an analogous situation, although he admitted to being incapable of devising how the details of such an experiment might go.

Even though it was originally proposed as a test for chimpanzees' mindreading capacities, Dennett's challenge was initially taken up by developmental psychologists Heinz Wimmer and Josef Perner (1983). They tested several three- to nine-year-old children using what was to remain for quite some time as one of the standard mindreading tests in developmental psychology—the 'false-belief task':

> In order to test subjects' comprehension of the other person's wrong belief, stories like the following were constructed: A story character, Maxi, puts chocolate into a cupboard x. In his absence his mother displaces the chocolate from x into cupboard y. Subjects have to indicate the box where Maxi will look for the chocolate when he returns. Only when they are able to represent Maxi's wrong belief ('Chocolate is in x') apart from what they themselves know to be the case ('Chocolate is in y') will they be able to point correctly to box x. This procedure tests whether subjects have an explicit and definite representation of the other's wrong belief. (Wimmer and Perner 1983, 106)

Wimmer and Perner found that none of the experimental subjects between three and four years of age could separate what they knew about the situation from what Maxi believed, and would incorrectly point to location *y* as the place where Maxi would go looking for his chocolate. In contrast, 57% of the four- to six-year-olds and 89% of the six- to nine-year-olds did point correctly to location *x*. This led Wimmer and Perner to conclude that mindreading capacities emerge and become fully developed at some point between four and six years of age. These results were widely replicated and, as this and other false-belief tests slowly became the gold standard for attributing mindreading capacities to an individual, the view that children under four cannot mindread became widely accepted (Andrews 2012b, 22-3). New experimental paradigms have, however, recently provided evidence of mindreading capacities in even younger humans (for a review, see Baillargeon, Scott, and He 2010), and so this common view is now being questioned by some.

Ironically, and despite its popularity in developmental psychology, it wasn't until 20 years after Dennett posed his challenge that a nonverbal false-belief task was finally designed and used to test a nonhuman species. Josep Call and Michael Tomasello (1999) were the ones to design this test. In the experiment, two identical containers were placed behind a barrier so that the experimental subjects could not see them. One experimenter, the hider, would hide a reward in one of the two containers. The experimental subjects could not see which one it was, as the barrier obstructed their view. Another experimenter, the communicator, was on the other side of the barrier and *could* see the container in which the reward was being hidden. In the training sessions, the communicator would place a marker on the correct container once the barrier had been removed so that the subjects knew which container to pick in order to get the reward. In the test condition, the communicator would leave the room after watching the hiding process. The hider would then switch the location of the containers. When the communicator returned, she would place the marker on the container located where she had originally seen the reward being hidden. The subjects could then choose the container that they thought had the reward in it. In order to choose correctly, according to the authors, the subjects had to be capable of understanding that "(1) the communicator is placing the marker on the container that is where she saw the reward being hidden, (2) the container that was there is now in the other location, and therefore (3) the reward must be in the container at the other location" (Call and Tomasello 1999, 382).

This non-verbal test was performed on chimpanzees, orang-utans, and human children. The latter were also subjected to a verbal version of the test, in which they could view the hiding and switching process, and were then asked to explain where they thought the communicator (who had not seen the switching) would place the sticker. Call and Tomasello obtained two very different results in this experiment. On the one hand, most of the children tested, who were between four and five years old, proved capable of passing the task, and there was also a strong correlation between their performance in the verbal and the nonverbal version of it, so that no child performed poorly on one and extremely well on the other (Call and Tomasello 1999, 387). The case of the nonhuman apes tested was quite different, however. Despite proving capable of passing the control tests, which showed that they were able to "track visible and invisible displacements, to remember the location of food, and to ignore the experimenter's marking behavior when the subject knew it to be wrong," none of the apes managed to pass the false-belief task itself (Call and Tomasello 1999, 391). Although absence of evidence is not equal to evidence of absence, as is often remarked in comparative psychology, this experiment's results failed to measure up to Dennett's challenge.

Call and Tomasello (1999)'s negative results were not alone at the time. The few experiments on chimpanzee mindreading that had been performed since Premack and Woodruff's original paper had all reached rather discouraging conclusions. These two authors themselves had performed another experiment to test chimpanzees' mindreading capacities (Woodruff and Premack 1979). In this test, the experimental subjects had visual access to the location of a piece of food but could not reach it. They could, however, let an uninformed experimenter know about the location. The uninformed experimenter could either be cooperative, and hand the food over to the chimpanzee once it was informed of its location, or she could be competitive, and decide to keep the food for herself. The experimenters wanted to test whether the chimpanzees would learn to deceive the competitive experimenter, which would provide evidence of their capacity to attribute intentions to others. They found that two out of four experimental subjects did end up providing misleading information to the competitor, but only after extensive familiarity with the test procedure, which suggests that they were behaving on the basis of an association, rather than engaging in mindreading (Call and Tomasello 1999, 382).

Similarly ambiguous results had been obtained in studies on chimpanzees' capacity to grasp the relation between seeing and knowing. In two separate experiments, David Premack (1988) and Daniel Povinelli, Kurt Nelson, and Sarah Boysen (1990) had managed to get chimpanzees to follow the indications of experimenters who *knew* where the food was because they had visual access to it, but not the ones of experimenters who were merely *guessing.* However, these positive results were obtained after several weeks of extensive training, and, so, were unsatisfactory because the chimpanzees had perhaps "simply learned a discrimination"

(Povinelli and Eddy 1996, 23).

The ambiguity in the results obtained up to that date was seemingly put to an end in 1996, when Daniel Povinelli and Timothy Eddy famously reported on a series of 15 different experimental studies they had performed in an attempt to determine whether chimpanzees could understand that others have visual experiences. In all these studies, the experimental subjects were placed in front of two experimenters: one who could see them (because her head was oriented in the right direction and she had unobstructed visual access to them), and one who could not (because she was facing in the wrong direction or because her vision was somehow obstructed). The aim of these studies was to determine whether chimpanzees would beg for food only from the experimenter who could *see* them, which would be an indicator that they can represent that others have perceptions—a type of mental state. Povinelli and Eddy found that, when one of the experimenters was facing forward and the other one was sitting with her back towards them, chimpanzees would only beg for food from the one who was facing forward. However, they would beg indiscriminately from both when one of them had something (such as a bucket over her head) obstructing her vision, or when she was not paying attention. This led Povinelli and Eddy to conclude that "even though young chimpanzee subjects spontaneously attend to and follow the visual gaze of others, they simultaneously appear oblivious to the attentional significance of that gaze" (Povinelli and Eddy 1996, vi), which would mean they do not understand seeing as a mental state.

In a later study, Daniel Povinelli *et al.* (1998) tested chimpanzees' understanding of intention by studying whether they could distinguish between intentional and accidental behaviour. The chimpanzees were subjected to two different tests. In one of them, they had to choose between receiving food from an actor that had previously spilled orange juice accidentally or from another actor who had intentionally spilled it. In the second test, they had to choose between receiving food from actors who either always succeeded in handing the food over to the chimpanzees, or always withheld the food themselves for their own consumption, or attempted to hand the food to the chimpanzees but were deprived of it by a third party. Insofar as intentions are mental states, had the chimpanzees always refrained from asking for food from the intentionally competitive actors, this test would have provided some evidence for mindreading capacities in them. However, "the subjects showed little evidence of using the accidental/inadvertent versus intentional distinction in their choices between the actors" (Povinelli et al. 1998, 205).

At the end of the 20th century, then, the prospects of finding convincing evidence of mindreading capacities in a nonhuman species were quite bleak. The turn of the century, however, saw a major change in the animal mindreading debate. Several studies had begun gathering evidence on chimpanzees' ability to follow the gaze of others around barriers (Tomasello, Hare, and Agnetta 1999), to check back to the one whose gaze they were following if they didn't see anything of interest (Call, Hare, and Tomasello 1998), and to eventually ignore the gazer if she repeatedly stared at something uninteresting (Tomasello, Hare, and Fogleman 2001). Despite these suggestive findings, the major advance in the chimpanzee mindreading debate came with the incorporation of a new experimental paradigm designed by Brian Hare *et al.* (2000). These experimenters' reasoning was that the tests performed until that moment had placed the chimpanzees in situations that were very unusual for them, as members of this species are "accustomed to competing for food, and indeed it is unlikely that they have ever experienced in their natural social lives a conspecific attempting to help them find food" (Hare et al. 2000, 772). Instead of expecting chimpanzees to beg for food from a human actor, the new tests now placed them "in the more natural situation of food competition with conspecifics," in hopes that this would trigger the use of their "most sophisticated social-cognitive abilities" (Hare et al. 2000, 783). Additionally, no previous training or human intervention was required in this novel test.

The test procedure went as follows. Two chimpanzees (a dominant and a subordinate) were placed in separate cages with a common arena in between. Food was placed in different locations within the arena and both chimpanzees were let into the room. The experimenters found that, when faced with a choice between food that the dominant conspecific could see and food that the dominant could not see, the subordinate individual would always go after the second choice. When a transparent barrier was placed between the dominant and the food, the subordinate would also refrain from going after that piece of food, and so she was apparently aware of the fact that a transparent barrier would not block the view of the dominant. Furthermore, whenever this same individual played the dominant role, she displayed the opposite strategy, going first for the food that was also visible to the subordinate, and only later retrieving the hidden one. These findings led Hare *et al.* to conclude that "chimpanzees know what conspecifics can and cannot see," in the sense that they can understand what they have visual access to and how this determines their behaviour, even though they are probably not aware of the fact that others' visual perception is analogous to their own (Hare et al. 2000, 784).

Further evidence of chimpanzees' understanding of visual perception was provided by Brian Hare, Josep Call, and Michael Tomasello (2006). In this experiment, chimpanzees competed against a human who would move food items out of their reach whenever they approached them in an attempt to retrieve them. They found that, when given the choice, the chimpanzees would prefer to approach the food from the side which a human actor's chest (but not her face) was oriented to, as well as from behind an opaque occluder rather than a transparent one. They would also choose the barrier that occluded their bodies fully rather than partially. The authors concluded that "chimpanzees can flexibly use knowledge of what a competitor can and cannot see to develop active, deceptive strategies for concealing their approach to contested food — and they do this from the very first trials in several novel situations" (Hare, Call, and Tomasello 2006, 508). Alicia Melis, Josep Call, and Michael Tomasello (2006) replicated these results, and also found that chimpanzees in such a situation would prefer to avoid a route to the food that triggered a loud noise. This suggests that chimpanzees may also have some understanding of others' auditory perception.

Josep Call *et al.* (2004) also provided some suggestive evidence of chimpanzees' understanding of others' intentions. They presented members of this species with a human that had food to hand over to them, but was either unwilling or unable to do so. The unwilling condition sometimes involved active teasing, such as waving the food in front of the chimpanzee but refusing to hand it over, and sometimes was a mere passive refusal. In the unable condition, there were obstacles such as physical barriers, or the human would feign distractedness or clumsiness. The chimpanzees' spontaneous behaviour was then observed and their reactions were found to be quite different. When the human was unwilling to hand over the food, the chimpanzees would produce more behaviours such as moving the apparatus or banging the panel that separated them from the experimenter. Additionally, when the human was unable to give them food, they would remain waiting for a longer period of time than when she was unwilling to do so. The authors concluded that these chimpanzees "did not just perceive others' behavior, they also interpreted it" (Call et al. 2004, 496).

In 2001, Brian Hare, Josep Call, and Michael Tomasello used the competitive paradigm to test chimpanzees' attribution of *knowledge* to others, purportedly a more difficult mental state to conceptualise than perceptions. In this experiment, the subordinate chimpanzee always had visual access to the location where a certain piece of food was being hidden (see figure 1). Additionally, from her location she could also see the dominant's enclosure and could potentially discriminate whether the latter was informed (i.e. had had visual access to the food's hiding process), uninformed (i.e. had had no visual access to it), or misinformed (i.e. the food was relocated after the dominant had seen the hiding process). The authors found that chimpanzees would go after the food more often when the dominant was either uninformed or misinformed of its location, as compared to when she was informed. Moreover, the subordinates would also choose to go after the food when the informed dominant was substituted by an uninformed one.

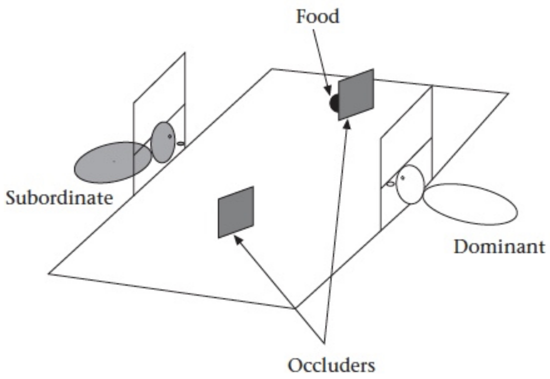

Figure 1: Experimental setup in Hare, Call, and Tomasello (2001). Source: Hare, Call and Tomasello (2001, 142).

A final experiment in this study got more ambiguous results. In this test, subordinates had visual access to the hiding process of two different pieces of food. They could also see that the dominant had witnessed one baiting, but was either uninformed or misinformed about the other. The experimenters expected the subordinates to go after the piece of food that the dominant was either uninformed or misinformed about, but they found no statistically significant trend in the subordinates' preference. The authors suggested that these results were probably due to a general change in the strategy followed by the dominant chimpanzees. These had begun to follow the subordinate to whichever location she approached, and so tracking what the dominant had witnessed was no longer relevant for the subordinates, for it had no effect on the former's behaviour (Hare, Call, and Tomasello 2001, 148). The authors concluded that, while this experiment provided evidence that chimpanzees "know what other individuals do and do not see" and "can recall what a conspecific has and has not seen in the immediate past," it is probably the case that they don't have as complex an understanding about others' visual perspective and how it relates to knowledge as humans do (Hare, Call, and Tomasello 2001, 149).

Juliane Kaminski, Josep Call, and Michael Tomasello (2008) also tested chimpanzees' understanding of others' states of knowledge/ignorance and false beliefs using a competitive paradigm. In the first test of this study, chimpanzees had to choose from several containers, only some of which had food in them. One of the subjects always knew where the food was, and had also seen the other chimpanzee either witnessing the baiting process or not, and so she could potentially modify her choice of container depending on whether her partner was knowledgeable or ignorant. The experimenters found that the subjects would act differently depending on whether they were choosing first or second. When they chose first, they would go after any piece of food, regardless of whether the competitor had seen it. When they chose second, however, they "tried to maximize their gains by guessing which cup the competitor had chosen" (Kaminski, Call, and Tomasello 2008, 229). In a second test, the subject witnessed the competitor being misinformed of the location of the food by an experimenter. When it was her turn to choose, she had to infer the location of the food by attributing a false belief to her competitor and thus predicting her incorrect choice of container. The chimpanzees showed an understanding of the knowledge/ignorance distinction by choosing the food that the competitor was uninformed of, but they showed no appreciation of others' false beliefs, as they did not prove capable of taking advantage of the misinformed condition. The authors concluded that chimpanzees are, at least in some situations, capable of understanding "when others are knowledgeable and ignorant — in the sense of what those others have and have not seen in the immediate past" but that they are incapable of distinguishing others' false beliefs from their true beliefs (Kaminski, Call, and Tomasello 2008, 233).

All of these studies, taken together, led Josep Call and Michael Tomasello to conclude, 30 years after David Premack and Guy Woodruff posed their original question, the following:

> In a broad construal of the phrase 'theory of mind', then, the answer to Premack and Woodruff's pregnant question of 30 years ago is a definite yes, chimpanzees do have a theory of mind. But chimpanzees probably do not understand others in terms of a fully human-like belief-desire psychology in which they appreciate that others have mental representations of the world that drive their actions even when those do not correspond to reality. And so in a more narrow definition of theory of mind as an understanding of false beliefs, the answer to Premack and Woodruff's question might be no, they do not. (Call and Tomasello 2008, 191)

They claim, then, that there is enough evidence to conclude that chimpanzees can mindread to a certain extent. That is, they have some understanding of others' perceptions and how they relate to states of knowledge and ignorance, as well as some comprehension of others' goals and intentions. The more complex, high-level forms of mindreading may, however, be restricted to humans. The answer to Premack and Woodruff's question, therefore, is a qualified *yes*, according to Call and Tomasello.

In addition to these studies on chimpanzees, a limited amount of evidence has been gathered that points to mindreading capacities in other species. There is, to begin with, some data concerning other, non-ape, primate species. Hika Kuroshima *et al.* (2002; 2003) found that capuchin monkeys could learn to follow the indications of a human actor who *knew* the location of food because she had *looked* inside the container where it was hidden, instead of the ones of an actor that was merely *guessing* where the food was. One of the experimental subjects learnt to generalise this discrimination to a variety of different containers and behaviours on behalf of the human actors, and even managed to distinguish the knower when the guesser was performing the same behaviours but directed at wrong locations. The authors took this as suggesting that at least this one monkey had managed to learn the relationship between seeing and knowing (Kuroshima et al. 2003, 289). Jonathan Flombaum and Laurie Santos (2005) found that rhesus monkeys spontaneously preferred to steal food from an experimenter who could not see them, as opposed to one who could, even if the only difference between them was the orientation of the eyes. Laurie Santos, Aaron Nissen, and Jonathan Ferrugia (2006) further found that members of this species would also prefer to steal food from a silent container as opposed to a noisy one when the experimenter was not looking, but would choose randomly when they were being watched. These two studies suggest that rhesus monkeys have some understanding of the relation between visual and auditory perception and knowledge.

Apart from these studies, there is a growing amount of evidence of mindreading capacities in corvids. Some species of this avian family will bury food items for later consumption (a behaviour known as 'caching') and will steal caches from conspecifics if faced with the opportunity. Nathan Emery and Nicola Clayton (2001) first discovered that scrub jays that had experience stealing food from conspecifics would re-cache their own food if there had been an observer present when the caching had originally occurred, but not if it had happened in private. Ravens were then found to cache food behind visual obstacles if an observer is present (Bugnyar and Kotrschal 2002), and to selectively recover food when a conspecific that knows its location, and thus poses a pilfering threat, is present (Bugnyar and Heinrich 2005).

Experiments with western scrub jays have also shown that they prefer to hide food in the shade, rather than in well-lit locations, if an observing conspecific is present (Dally, Emery, and Clayton 2004), that they will also choose distant or out-of-view locations for their caches when observed, and, when in private, will selectively recover those caches whose location the observer has had visual access to (Dally, Emery, and Clayton 2005; Emery, Dally, and Clayton 2004). It has further been found that they can distinguish knowledgeable from ignorant conspecifics and alter their behaviour accordingly (Dally, Emery, and Clayton 2006).

All this evidence seems to suggest, then, that some species of monkeys and some species of corvids, together with chimpanzees and perhaps other great apes, understand the relationship between perceiving and knowing, at least in a rudimentary fashion. In the case of chimpanzees and perhaps corvids, the data suggests that they may also understand that others possess intentions or goals. It is important to bear in mind, however, that it is only fairly recently that interesting results have begun to be obtained in the animal mindreading studies, and very few species have yet been tested. It may thus be the case that more evidence will be gathered in upcoming years. Nevertheless, some authors consider that none of the data reviewed constitutes real evidence of mindreading capacities in a nonhuman species, and that there are no reasons to believe that any other animal, besides us, can attribute mental states to others. In the following section, we shall have a look at these authors' reasons.

3.3. THE BEHAVIOUR READING HYPOTHESIS

In the previous section, we saw that many of the researchers and scholars taking part in the animal mindreading debate would agree that the evidence shows that at least *some* nonhuman species have *some* understanding of others' mental states. A small number of scholars, however, remain entirely sceptical with respect to the possibility of animal mindreading. This scepticism is most commonly associated with Daniel Povinelli and colleagues (Povinelli and Vonk 2003; Povinelli and Vonk 2004; Penn and Povinelli 2007; Penn and Povinelli 2013; Penn, Holyoak, and Povinelli 2008), but has also been manifested by Cecilia Heyes (1998; 2014), and Robert Lurz (2009; 2011). This section will be devoted to the reasons behind these scholars' position.

The sceptics argue that the data gathered by the different experimenters doesn't even come close to *suggesting* that animals possess mindreading capacities, and that this is not due to the kind of results obtained, but rather, to a *fundamental flaw* in the experimental paradigms used. In 1998, Cecilia Heyes first argued that the procedures used in the experiments performed until that moment didn't have "the potential to yield evidence favoring a theory of mind interpretation over other current candidates" (Heyes 1998, 102). Even though they can inform us of *which* observable cues the animals use when deciding how to interact with others, she argues, they cannot tell us *whether* these cues are being used because, in the animal's understanding, they signify certain mental states (Heyes 1998, 108). The experimental paradigms she refers to (e.g. Povinelli and Eddy 1996) were the ones used before Brian Hare and colleagues introduced the competitive paradigm. The reader may recall that Hare's was presented as a much more ecologically valid paradigm, and so presumably it is better suited to determine an animal's mindreading capacities than previous experimental designs. However, in 2004, Daniel Povinelli and Jennifer Vonk wrote that even these experiments "do not, in principle, have the ability to provide evidence that uniquely supports [the mindreading] hypothesis" (Povinelli and Vonk 2004, 2). Several years later, Robert Lurz defended that "all current mindreading tests with animals, by virtue of their very design or 'logic,' are liable to produce false positives" (Lurz 2011, 23). And still in 2013, Derek Penn and Daniel Povinelli asserted that the debate suffered from "a lack of compelling evidence for anything remotely resembling [mindreading capacities] among nonhuman animals" (Penn and Povinelli 2013, 69).

The problem that these scholars are talking about, that is, the fundamental flaw in the animal mindreading research paradigms, is what Robert Lurz (2009; 2011), following Susan Hurley and Matthew Nudds (2006), calls the 'logical problem,' and Cecilia Heyes (2014) calls the 'observables problem.' The argument goes as follows:

> P_1: Others' mental states cannot be subjected to direct observation.
>
> P_2: All mental state attribution must be performed on the basis of behavioural or environmental cues.
>
> P_3: Current mindreading tests can all be passed both by reading these behavioural and environmental cues in a certain manner, and by attributing mental states to others.
>
> P_4: None of the current mindreading tests has the capacity, even in principle, to distinguish a mindreading response from its behaviour reading counterpart.
>
> C: Therefore, there is no evidence in favour of the hypothesis that nonhuman animals can mindread.

What is being presented as an alternative to the mindreading explanation of the experimental results, is what has been called the *behaviour reading hypothesis*. This hypothesis was originally suggested, somewhat implicitly, by Cecilia Heyes (1998), and was fully developed by Daniel Povinelli and colleagues (Povinelli and Vonk 2003; Povinelli and Vonk 2004; Penn and Povinelli 2007; Penn and Povinelli 2013; Penn, Holyoak, and Povinelli 2008), as well as by Robert Lurz (2009; 2011). The idea, then, is that those animals that pass the different mindreading tests may not, after all, be mindreaders, but simply *behaviour readers*.

Behaviour reading has been constructed as an alternative to mindreading that does not imply committing oneself to the claim that animals are mere *behaviouristic* beings. If behaviourism were to offer an accurate description of how animals function in the world, then these would only be capable of behaving "on the basis of brute associations" (Lurz 2011, 25), and lack "any ability to reason about the causal relation between [others'] behaviors in an abstract or inferentially coherent fashion" (Penn and Povinelli 2013, 62). The behaviour reading hypothesis does not require such low-level mechanisms to be at the basis of all animal behaviour. Rather, it is fully compatible with the idea of certain animals as "cognitive creatures capable of reasoning about general classes of past and occurrent behaviours" (Penn and Povinelli 2007, 733), that can "employ both inferential and simulative mechanisms for forming abstractions about classes of behaviours and environmental conditions that are relevant to their goal-directed actions," and "are able to generalize the lessons learned from these abstractions to novel scenarios" (Penn and Povinelli 2007, 737). Thus, to borrow Povinelli and Vonk's example, a behaviour reader does not need to represent and understand separately the different forms that threat behaviour may take. Rather, she can group these different manifestations as instances of the more general and abstract category of 'threat behaviour' (Povinelli and Vonk 2003, 157). This would be what allows dogs, for instance, to expect an attack when they see a conspecific growling and also when they see that she is bearing her teeth. Being able to employ abstract categories would allow behaviour readers to respond adequately and flexibly to others' behaviour. And importantly, they would acquire this without the intervention of mindreading capacities. Indeed, the core idea of the behaviour reading hypothesis is the following:

> [A]nimals *lack mental state concepts* and predict or respond to another agent's actions by knowing (either from past experience, or by inference, or perhaps even innately) that certain observable features of the other agent or its environment are reliable indicators of its future behavior. (Lurz 2011, 28, my emphasis)

The behaviour reading hypothesis allows us to reinterpret, in a non-mentalistic fashion, the animals' performance in the different mindreading tests. For instance, the original cooperative-begging paradigm that was used to test chimpanzees' attribution of visual perception to others (Povinelli and Eddy 1996) could be passed if the chimpanzees were able to reason, perhaps on the basis of past experiences, that experimenters are more likely to respond to their begging gestures when there is a direct line of gaze between the experimenters' eyes and the chimpanzees themselves. It would not have been necessary for the chimpanzees to reason, additionally, that this is due to the fact that, when there is such a direct line of gaze, then it means that the experimenters can *see* that they are begging for food (Heyes 1998, 108). This sort of reinterpretation is also possible for the results obtained in the competitive research paradigm introduced in Hare *et al.* (2000). The subordinates could pass this test even if they were incapable of understanding that the dominant can *see* the food that is behind the transparent barrier, and *not see* the one that is behind the opaque barrier. All that is required is for them to follow a behavioural rule such as 'do not approach food that a dominant has a direct line of gaze to,' where 'direct-line-of-gaze' is understood as something that does not obtain in the presence of an opaque barrier (Lurz 2011, 33—5). The same applies for all attribution of knowledge/ignorance tests based on visual perception, as Lurz notes:

> Since judgments of whether an agent *saw/did not see* an event rest upon judgments of whether the agent *had/lacked* a direct line of gaze to the event, all knowledge/ignorance attribution studies are amenable to a complementary behavior-reading explanation in terms of the animal predicting another agent's behavior by understanding what the agent had or did not have a direct line of gaze to in the past. (Lurz 2011, 50, his emphasis)

Similarly, in the different tests with corvids, it is only necessary for the birds to "keep track of 'who' was present for which caching event without, *in addition*, keeping track of the distinct counterfactual representations being maintained by each individual competitor" (Penn and Povinelli 2013, 70, their emphasis). The same sort of analysis can be applied to tests of goal attribution, where goals need only be understood in an external manner, as an outcome towards which the other's behaviour is oriented, without any understanding of how the other agent represents it (Penn and Povinelli 2013, 71). The crucial idea is that the animals' performance in these tests would be *the same*, regardless of whether they were behaving on the basis of mentalistic reasoning, or on the basis of a non-mentalistic interpretation of behavioural and environmental cues. There is, these authors argue, no way in which these tests can help us determine the answer to the question of whether any other animal, besides humans, can attribute mental states to others. Accordingly, these sceptics have devoted much of their efforts to devising new, alternative experimental paradigms that can help us address this question unambiguously.

Heyes proposed an experimental design in which chimpanzees would be first allowed to manipulate and familiarise themselves with two different goggles, indistinguishable from the outside except for the colour of their rims, but one of which allowed the subject wearing them to see, and the other not. In the test phase, the chimpanzee would be placed in front of two experimenters, each of whom would be wearing one of the goggles. The chimpanzee would be deprived of visual access to the hiding process of a piece of food in one of two containers. Both experimenters would be present when the baiting took place, but only the one wearing the translucent goggles would be capable of seeing where the food was hidden. The test would then consist of determining whether the chimpanzee would follow the cue of the experimenter who had seen the baiting process, but not the other's. If the chimpanzee passed the test, according to Heyes, this would mean that she would have been able to infer that, since only one of the two goggles had allowed herself to *see*, she should beg from the experimenter wearing this pair of goggles, as she will be the one that will have *seen* the baiting process. Povinelli and colleagues proposed a version of this test, using the cooperative-begging paradigm. The experiment would test whether chimpanzees would request food from the experimenter wearing the see-through device, as opposed to the one wearing the opaque one (Povinelli and Vonk 2003; Povinelli and Vonk 2004; Penn and Povinelli 2007).

Lurz (2009; 2011), however, has argued that neither Heyes' nor Povinelli *et al.*'s protocols could solve the logical problem, for the chimpanzees could pass the test simply by realising that the opaque visors prevent themselves from having a direct line of gaze to certain objects, and so that this will be the case too for the experimenters wearing them. He suggests that the way to solve the logical problem, at least for attributions of visual perception, is by using protocols where the animals would have to attribute seeing-as to other individuals in order to pass the test. For instance, in one of his proposed experiments, a subordinate and a dominant chimpanzee would be placed in separate compartments with an arena in between, like in Hare's competitive paradigm. Both chimpanzees would be familiarised with fake orange bananas made of plastic but indistinguishable at a distance from regular bananas except for their colour. The subordinate would then be given the chance to explore how different objects look behind transparent and red-coloured translucent screens. In the test condition, the fake banana would be placed behind the clear screen and the regular banana would be placed behind the red screen. Both chimpanzees would then be let into the arena. If the subordinate chimpanzee decided to go after the regular banana, then we could conclude that she probably has the capacity to understand that the dominant *sees it as* orange (because it is behind the red screen), and thus thinks that it is a fake banana. Despite the ingenuity of this protocol, some authors (Andrews 2012a; Buckner 2014) have argued that it doesn't have the capacity to solve the logical problem either, for it can also be passed solely by reasoning on the basis of behavioural and environmental cues. For instance, Andrews has argued that "the subordinate could understand the clear transparent barriers alone as offering a direct line of gaze, and the red transparent barriers as strange blockers or modifiers of direct line of gaze," which means that the subordinate could presumably "predict that the dominant would move toward the banana to which the dominant has a direct line of gaze" without having to think about "how the yellow banana appears to the dominant" (Andrews 2012a).

Recently, Thomas Bugnyar, Stephan Reber, and Cameron Buckner (2016) published a study where they claimed to have solved the 'logical problem.' In this study, performed on ravens, the experimental subjects were placed in a room that was separated from an adjacent room by a wooden wall. This wall had two windows that allowed visual access from one room to the other, and could be blocked by a cover. Each window had a peephole attached that could be opened or closed independently, and which the ravens were not familiar with before this study. The experiment was then performed in three steps. In the first phase, the peepholes were kept shut and the ravens were allowed to cache food with a conspecific in the adjacent room and one of the windows open or closed. In the second phase, the subjects were trained to use one of the peepholes in order to observe where a human experimenter was hiding food that they could later pilfer. In the final phase (see figure 2), the ravens were subjected to two baseline conditions, in which there was a conspecific in the adjacent room and the windows were either open or closed, and a test condition, in which both windows were closed and the peephole was open. In the test condition there was no competitor in the adjacent room, but the subjects could hear recordings of the competitor that had been taken during the non-observed condition. The ravens treated the test condition as equal to the observed condition, by finishing their caches faster and returning to improve them less often than during the non-observed condition. The authors interpreted this as a sign that they can "generalize from their own experience using the peephole as a pilferer and predict that audible competitors could potentially see their caches" (Bugnyar, Reber, and Buckner 2016, 3). The authors further argued that "all available behaviour-reading cues [had] been controlled for in the test condition," given that "there is no actual competitor whose gaze could be read, and the situation is novel from the subject's perspective." This led them to conclude that these results constitute "clear evidence that raven social cognition cannot be reduced to behaviour-reading" (Bugnyar, Reber, and Buckner 2016, 3).

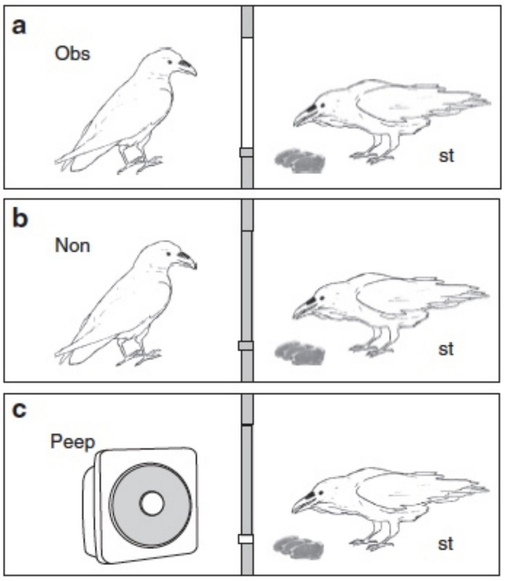

Figure 2: Experimental setup in Bugnyar, Reber, and Buckner (2016). Boxes (a) and (b) represent the baseline conditions, and (c) represents the test condition. Source: Bugnyar, Reber, and Buckner (2016, 2).

While this study certainly constitutes the most suggestive evidence to date, it could be argued that its results could also be reinterpreted along the lines of the behaviour reading hypothesis. During the training phase, the ravens had perhaps learnt to classify the peephole as pertaining to the same class of environmental stimuli as the window, that is, as a barrier that, when open, affords the possibility of pilfering others' caches. This, coupled with the fact that they probably interpreted the recordings in the test condition as a sign that there was a conspecific in the adjacent room, would have led them to behave in a way that, as they had perhaps learned from previous experiences, minimises the risk of their caches being pilfered. While a reasoning in terms of others' perceptions seems to us, skilled mindreaders as we are, the most straightforward way of understanding the relation between the absence of a visual occluder and a threat to our property, it is at least possible that the ravens simply understood the peephole in affordance-based terms as something that, when open, grants the possibility of pilfering others' caches, and when closed, doesn't. If this were correct, it would appear that Bugnyar, Reber, and Buckner haven't, after all, managed to solve the 'logical problem,' which would leave us with the question of whether *any* experimental setup can provide a complete and satisfactory solution to this problem.

The behaviour reading hypothesis, as we've seen, is fundamentally a critique of the current experimental paradigms in the animal mindreading debate, but Povinelli and colleagues sometimes seem to adhere to a stronger claim, namely, that *no animal, except humans, can mindread*. For instance, they assert that "considerable evidence suggests that [animals] are not making such [mentalistic] inferences in situations where humans (children and adults) would readily do so" (Povinelli and Vonk 2004, 21), that the possibility that animals may be mindreaders is as "implausible" as the possibility that they may be mere behaviouristic beings (Penn and Povinelli 2013, 63), and that the capacity to reason "in a symbolic-relational fashion," which, they argue, is necessary for mindreading, "evolved in only one lineage — ours" (Penn, Holyoak, and Povinelli 2008, 129). We can, therefore, distinguish two different versions of the behaviour reading hypothesis:

> *Behaviour reading hypothesis — weak version*: All current experimental results in the animal mindreading debate may be reinterpreted in a non-mentalistic fashion, and, so,

there is no evidence that animals can mindread. This claim is explicitly endorsed by Heyes, Lurz, and Povinelli *et al.*

Behaviour reading hypothesis — strong version: Some nonhuman animals are behaviour readers, but only humans are mindreaders. This claim is implicitly endorsed by Povinelli *et al*.

The main target of this dissertation is to argue that mindreading is not required for moral subjecthood, and thus that the animal mindreading debate and the animal morality debate should be understood as independent from each other. In order to make my case, I am going to assume the worst-case scenario. That is, I'm going to assume that the strong version of the behaviour reading hypothesis is true and that no other animal, besides humans, can engage in mindreading. I shall assume that the animals that are possible candidates for moral subjecthood (great apes, some cetaceans, elephants, some corvids, dogs, and so on) can be nothing more than behaviour readers, and, so, that they can, at most, "(a) construct abstract categories of behavior, (b) make predictions about future behaviors that follow from past behaviors, and (c) adjust their own behavior accordingly" (Povinelli and Vonk 2003, 157), but lack all capacity to attribute mental states to others. While the behaviour reading hypothesis, in both its versions, has been criticised by some scholars, who have accused it of being empirically unsolvable (Andrews 2005; Heyes 2014), unparsimonious (Call and Tomasello 2008), and lacking semantic clarity (Buckner 2014), assuming the truth of the strong version of this hypothesis serves my purpose, which is to argue that we may find moral subjects in the animal kingdom, even if mindreading is an exclusively human capacity. The upcoming chapters of this dissertation will thus be concerned with arguing that *the lack of mindreading capacities is not an obstacle for the possession of moral subjecthood*.

3.4. SUMMARY

This chapter has offered an overview of the debate on animal mindreading. Despite having gone on for almost 40 years, researchers and scholars are still far from a consensus on which nonhuman species, if any, possess mindreading capacities. We have seen that a number of scholars, most notably Josep Call, Michael Tomasello, and Brian Hare, consider that there is enough evidence to attribute some mindreading capacities (namely, an understanding of others' goals or intentions, visual and auditory perceptual states, and how these relate to states of knowledge/ignorance) to great apes, and some species of monkeys and corvids. The presence of a more complex, metarepresentational form of mindreading, such as is necessary for the attribution of false beliefs to others, these scholars argue, is probably an exclusively human capacity.

At the other side of the spectrum, sceptics such as Daniel Povinelli, Robert Lurz, and Cecilia Heyes, have argued that the experimental protocols used to determine whether animals can mindread suffer from a fundamental flaw that renders them unable to discriminate a mindreading response on behalf of the experimental subjects from its behaviour reading counterpart. Therefore, these authors argue, there is no evidence that animals can mindread. Additionally, Daniel Povinelli and colleagues seem to adhere to a stronger claim, namely, that *only* humans can mindread. In the following chapters, I will assume that this stronger claim is true, and will argue that behaviour readers may nevertheless be moral subjects.

This chapter has also been concerned with sketching the theoretical reasons behind the pervasive assumption that morality is exclusively reserved for mindreaders. We have seen that these are of two distinct sorts. On the one hand, moral psychologists assert, from an *a posteriori* standpoint, that mindreading is present in moral judgement, because our moral judgements *de* *facto* engage our mindreading capacities. On the other hand, moral philosophers tend to assume, from a largely *a priori* standpoint, that mindreading is necessary for moral behaviour. It is this second claim that I am interested in arguing against. I have shown that, apart from the role of mindreading in a critical scrutiny of our motivations, there are four main conditions that have been defended as necessary for the presence of moral behaviour, each of which would separately entail that it requires mindreading capacities. These conditions appeal to the *reciprocal* and *caring* nature of moral behaviour, as well as to the need for some sort of understanding of others' mental welfare in order to be properly *motivated* to care for them and to be *responsible* for our acts.

The possibility that animals may be moral would be rendered largely implausible if these authors were right and morality were indeed dependent upon mindreading. As we saw in the previous chapter, there is a very large, and growing, set of evidence that points to the presence of apparently moral behaviour in quite a big number of species, including great apes, monkeys, and corvids, but also canids, elephants, cetaceans, rodents, and perhaps further species. The case for animal morality would be dealt a severe blow if it were indeed dependent upon the case for animal mindreading, as it seems unlikely that this many species will possess an understanding of others' mental lives. In the following chapters, I will try to show that these two debates are completely independent from each other, and furthermore, that the existence of moral behaviour readers is theoretically conceivable, and perhaps even empirically plausible.

4. EMPATHY AND MINDREADING

I am going to argue that there can be moral motivations that do not require the possession of mindreading capacities. We saw in the previous chapter that this is a claim that will go against the pervasive assumption, presupposed by philosophers and scientists alike, that morality requires mindreading. My interest is not in arguing that this assumption is an absurdity—on the contrary, I think it's quite plausible that many, if not most, instances of moral motivation and moral behaviour in humans involve, either explicitly or implicitly, some mindreading component. Indeed, it seems safe to assume that our capacity to understand each other as individuals with mental lives greatly shapes our interactions. I want to argue that, in spite of this, there is *at least one* type of moral motivation that does not require any mindreading component *at all*, that is, that can be possessed by individuals that are mere *behaviour readers*. The type of moral motivation in question will be *moral emotions*.

To defend the coherence of the notion of moral emotions in behaviour readers, I am going to focus on a specific example—the case of empathy—and then briefly sketch how the analysis I shall put forward, which is greatly influenced by Mark Rowlands (2012b), can be extrapolated to account for other moral emotions. In chapter 5, I will argue that it is conceivable, and perhaps even empirically plausible, that minimal forms of empathy that suffice to count as a moral emotion act as a motivation for certain animals' behaviour, and that this may be the case even if no nonhuman animal is capable of understanding others as possessors of mental lives. Before describing the form of empathy that may serve as a moral motivation in behaviour readers, it is important to look at how the term 'empathy' is used in the literature, and what relations are thought to bear between this notion and that of mindreading. The term ‘empathy,’ as we shall see, is notoriously ambiguous, which is why I will devote this chapter to disentangling its different meanings.

In recent years, there has been an upsurge of interest in the notion of empathy coming from several different disciplines, ranging from social neuroscience and philosophy of mind, to ethics, moral psychology, and developmental psychology. Much of the effort on behalf of theorists from these different areas has been devoted to definitional issues. As a result, there is now a vast amount of literature dedicated to the topic of how the notion of empathy should be characterised, but there is still little consensus on what this word actually denotes. We have reached a situation in which “[t]here are probably nearly as many definitions of empathy as people working on the topic” (de Vignemont and Singer 2006, 435).

Daniel Batson (2009) argues that there are eight distinct phenomena that the word ‘empathy’ can be used to refer to, and points out that this disagreement amongst theorists partly stems from the fact that said notion is used to answer two very different (albeit partially related) questions:

> (a) “How can one know what another person is thinking and feeling?,” and
>
> (b) “What leads one person to respond with sensitivity and care to the suffering of another?” (Batson 2009, 3)

Each of these questions lies at the basis of a different research strand on the topic of empathy. In the first of these strands, which corresponds to question (a), empathy is understood as a mechanism related to interpersonal understanding, but this is as far as the consensus goes, for exactly what empathy is and how it contributes to social cognition is a matter of much debate (see, e.g. Michael 2014). In the second of these research strands, which corresponds to question (b), empathy is understood as a motivation for prosocial/altruistic/moral behaviour and here, once again, there are many different ways in which the term ‘empathy’ is used (see, e.g. Preston and de Waal 2002; de Waal 2008). To make matters more complicated, there are some characterisations of the notion of empathy that, as we shall see, stand somewhere in between both of these research strands, with empathy being understood as a form of interpersonal understanding that can, at the same time, lead to prosocial/altruistic/moral behaviour. Moreover, some authors adopt a multi-level conception of empathy, thus subscribing different characterisations of the term simultaneously.

The literature on empathy is vast, and it's not necessary for my purposes to attempt a comprehensive review. However, it is rather important, in order to avoid confusions in the following chapter, to map out the different ways in which this term is used, and the sort of relationship that is thought to bear between empathy, morality, and mindreading in the different strands of the debate. This chapter will be divided into three parts. In the first part, I will focus on notions of empathy that are linked to social cognition. In the second part, I will analyse the different ways in which empathy is thought to play a role in moral motivation. In the last part, I will go back to the empirical literature on animal morality and present a common interpretation of the evidence from the perspective of the empathy debates. My focus, throughout this chapter, will be to circle out the notions of empathy that will be relevant for my own argument and the specific claims within these debates that I will attempt to refute in the following chapter.

4.1. EMPATHY AS SOCIAL COGNITION

One of the aims of debates on social cognition is to give an account of how social beings, such as ourselves, come to know how others think and feel. If we understand the term 'mindreading,' as I've been doing, in its broadest possible sense—that is, as the conceptualisation and attribution of mental states to others—, then we can say that these debates attempt, at least partially, to give an account of how social beings come to mindread. By understanding the term ‘mindreading’ in this broad sense, I intend to let go of any presuppositions about how this ability is actually implemented in species like our own. Accordingly, I will understand that authors, such as Shaun Gallagher and Daniel Hutto (e.g. Gallagher 2008; Hutto and Gallagher 2008), who defend that we come to know how others think and feel by direct perception and/or interaction, instead of through inference or simulation, are also attempting to give an account of our mindreading abilities. I will discuss this in more detail throughout this section.

The different notions of empathy that can be found within debates on social cognition have their origin in the German term ‘Einfühlung’ (literally, ‘feeling into’), which the psychologist Edward Titchener first translated as ‘empathy’ at the beginning of the 20th century (Titchener 1909). The term ‘Einfühlung’ appeared in the second half of the 18th century as a way of describing the aesthetic experience of ‘feeling into’ works of art or natural objects and grasping their beauty. By the end of the 19th century, it had become an important concept in German philosophical aesthetics. Theodor Lipps (1903) was the first to use this term in a broader sense to refer to the ability that allows us, not only to have an aesthetic appreciation of different objects, but also to perceive those around us as minded creatures. This is the conception of ‘Einfühlung’ that Titchener had in mind when he translated it as ‘empathy’ (Stueber 2014). Current debates on empathy and social cognition can be traced back to this first use of the term, since ‘empathy’ is here understood as a means for interpersonal understanding, and not as a moral motivation. In contrast, the use of the term ‘empathy’ to refer to a moral motivation comes closer to the notion of sympathy that appears in the works of David Hume and Adam Smith (see section 4.2).

Despite having a common conceptual origin, many disparate notions of empathy can be found within debates on social cognition. By looking at how different authors that engage in these debates characterise the relation between empathy and mindreading, we can distinguish three main ways in which the term ‘empathy’ is used. In a first sense, which is more often associated with the verb ‘to empathise’ rather than the noun ‘empathy,’ empathising is barely distinguishable from mindreading, that is, from accessing what others think and feel. The two verbs are used interchangeably, thus becoming synonyms. Following John Michael (2014, 157), we can distinguish two further ways of understanding the relationship between empathy and mindreading within this research strand. On the one hand, there is the notion of empathy as an ability or mechanism that enables mindreading. On the other hand, there is the notion of empathy as an ability or mechanism that presupposes an ability to mindread, but is no longer a synonym of this term, like in the first case, but rather something more. I will now go on to discuss these three uses of the term 'empathy.'

4.1.1. Empathy as mindreading

Let us begin by looking at those uses of the term 'empathy' that equate it to the term 'mindreading.' Here, empathy is understood as the effective attribution of mental states to another. Empathising with someone thus means accessing what she thinks or what she feels. Within debates on social cognition, this form of empathy is an *explanandum*, that is, one of the phenomena that researchers seek to explain (Batson 2009, 8). While we can find several examples of this use of the term 'empathy' in the literature, there are several subtle, and some not so subtle, differences in how this concept appears in different texts. It is not easy to find several different authors that possess one and the same understanding of empathy as (roughly) mindreading. The situation is best described as a cluster of different uses of the noun 'empathy,' or the verb 'to empathise,' all of which resemble the notion of mindreading.

A clear use of the term 'empathising' as a synonym for 'mindreading' can be found in a 1998 issue of *Scientific American*, in which Gordon Gallup, Jr. and Daniel Povinelli contributed with two separate articles, each of which attempted to answer the question "Can animals empathise?" Even though no explicit definition of empathy was endorsed by either of these two authors, the way they both phrased their answers quite clearly suggests that they understood empathising to be equivalent to mindreading. First, Gallup answers the question of whether animals can empathise by saying "yes," and giving the explanation that "[a]nimals that pass the mirror test are self-aware and thus can infer the states of mind of another individual" (Gallup Jr. 1998, 66). This suggests that he considers empathising to mean inferring the mental states of others, which is a form of mindreading. Daniel Povinelli also seems to understand empathy in this same sense, for he answers the question of whether animals can empathise by saying "maybe not," and asserting that "[e]ven though chimpanzees pass the mirror test, they do not seem to conceive of others' —or even their own— mental states" (Povinelli 1998, 67), which suggests that he considers empathising to mean, precisely, conceiving of others' mental states. He also claims that "one of the most basic empathic aspects of human intelligence [is] the understanding that others *see*" (Povinelli 1998, 73, his emphasis). Since perceptions are a type of mental state, an understanding of another's perceptions is a form of mindreading. These two articles thus appear as a clear example of 'empathising' being used as a synonym for 'mindreading.'

A more ambiguous use of the term 'empathy' is found in an article written in 1986 by Lauren Wispé. In it, she argues that empathy could not be experimentally demonstrated without first getting rid of the ambiguity associated with this term (Wispé 1986, 317). To this end, she proposes that the term 'empathy' be reserved to refer to "the attempt by one self-aware self to comprehend unjudgmentally the positive and negative experiences of another self" (Wispé 1986, 318). Non-judgementally comprehending another individual's positive and negative experiences, insofar as these are mental states, would certainly be a form of mindreading, and so it seems that Wispé is thinking of empathising as essentially mindreading. However, there is a certain ambiguity in her definition (brought by the use of the term 'attempt') that suggests that Wispé may be thinking of empathy as a mechanism that *enables* mindreading (see section 4.1.2.). Her subsequent clarifications don't succeed in eliminating this ambiguity. On the one hand, she states that when we empathise "we substitute ourselves for the others," which suggests that empathy is a process that enables the attribution of mental states to others. On the other hand, however, she also states that "[t]o know what it would be like if *I* were the other person is empathy" (Wispé 1986, 318, her emphasis), which brings it closer to an understanding of empathy as mindreading itself. The true meaning behind Wispé's words is, luckily, not important for our purposes.

One of the most widely cited papers on empathy is Stephanie Preston and Frans de Waal's (2002), which was published in *Behavioral and Brain Sciences* with an open peer commentary. In this paper, Preston and de Waal attempted to construct a comprehensive account of the ultimate and proximate bases of empathy that would integrate into one mechanism all the different mental phenomena that this term can be used to denote. This means that they simultaneously endorsed many different notions of empathy. In particular, what they term "cognitive empathy" seems to closely resemble the notion of empathy we are currently considering, for they characterise it as the process whereby a "[s]ubject represents [the] state of [the] object through top-down processes" (Preston and de Waal 2002, 4). This definition is ambiguous insofar as "state" can be taken to mean very different things, and not necessarily mental states. However, Preston and de Waal's posterior discussion strongly suggests that they take "state" to mean, at least partially, mental state (see: Preston and de Waal 2002, 18-20), which brings the notion of cognitive empathy close to that of mindreading. On the other hand, they also seem to understand "cognitive empathy" as a form of prosocial motivation (see: Preston and de Waal 2002, 19). This doesn't necessarily have to be inconsistent with the notion of cognitive empathy as mindreading—knowing that you are in pain may, for instance, be a perfectly valid motivation for me to comfort you. Some of the commentaries that appeared in this issue of *Behavioral and Brain Sciences* did fully and explicitly endorse the notion of cognitive empathy as a synonym for mindreading. James Blair and Karina Perschardt referred to "cognitive empathy" as both "Theory of Mind" and "the ability to represent the mental states of others" (Blair and Perschardt 2002, 27). Gordon Gallup, Jr. and Steven Platek also endorsed this notion, by claiming that "[t]he subcategory of empathy that Preston & de Waal ... identify as cognitive empathy represents an instance of a more general phenomenon known as mental state attribution" (Gallup Jr. and Platek 2002, 36).

Another example of this notion of empathy can be found in Paul Eslinger (1998), who states that empathy can be characterised as "the 'understanding' of another's experiences and emotional states, dependent upon cognitive processes" (Eslinger 1998, 194). The definition of mindreading I have endorsed, as the conceptualisation and attribution of mental states to others, would certainly imply an understanding of said mental states and be dependent upon cognitive processes, thus closely resembling this notion of empathy. A similar idea to Eslinger's is found in William Ickes (2009), who uses the term "empathic inference" to refer to "the everyday mind reading that people do whenever they attempt to infer other people's thoughts and feelings," and the term "empathic accuracy" to denote "the extent to which such everyday mind reading attempts are successful" (Ickes 2009, 57).

Dan Zahavi has defended a notion of empathy that is also, at least arguably, an example of empathy as mindreading (see, e.g.: Zahavi and Overgaard 2012; Zahavi 2014). He subscribes a characterisation of empathy that is greatly influenced by the phenomenological tradition. While he explicitly denies that empathy is "a question of just abstractly ascribing a certain mental state to another" (Zahavi 2014, 138), he clearly considers it as a form of access to others' mental states. He writes:

> [Empathy] is a distinctive form of other-directed intentionality, distinct from both self-awareness and ordinary object-intentionality, which allows foreign experiences to disclose themselves as foreign rather than as own ... We can see the other's elation or doubt, surprise or attentiveness in his or her face, we can hear the other's trepidation, impatience or bewilderment in her voice, feel the other's enthusiasm in his handshake, grasp his mood in his posture, and see her determination and persistence in her actions ... Empathy is the experience of the embodied mind of the other (Zahavi 2014, 138)

Zahavi points out that the sort of access to another's mental states that is provided by empathy is "basic and intuitive," which makes empathic understanding of others very different from those forms of understanding that are obtained via theoretical inference or simulation. While taking care to explicitly distinguish his notion of empathy from mindreading as characterised by theory-theorists and simulation theorists, Zahavi characterises empathy as "a perception-based experiential access to the minds of others" (Zahavi 2014, 139). Even if Zahavi does not understand others' minds as completely hidden from perception, but as something that can be experienced to a certain extent, it seems reasonable to classify his notion of empathy as a form of mindreading, only a direct and perception-based one (more on this in section 4.1.3 below).

There are subtle differences between these different definitions of empathy, and so they may not all refer to exactly the same phenomenon. The phenomena they cover are, however, similar enough to constitute a class that exemplifies one particular tendency in definitional debates on empathy, namely, the one that brings it close to the notion of mindreading. This does *not* constitute a form of empathy that I'm in any way interested in for present purposes, as should be clear if we take into account that I will be attempting to give a characterisation of empathy that will make it entirely *independent* of all capacity to mindread. In upcoming chapters I will therefore leave this notion of empathy completely aside.

4.1.2. EMPATHY AS ENABLING MINDREADING

As we saw in the previous section, the notion of empathy as a synonym for mindreading often appears linked to conceptions of mindreading that are either inference-based or perception-based. The notion of empathy as *enabling* mindreading, in turn, can frequently be found in authors who have an understanding of mindreading that is simulation-based. Here, empathy no longer refers to the very act of becoming acquainted with others' mental states, but rather, to the process, mechanism, or ability that allows us to have indirect knowledge of another's mental life. The idea is that the very act of mindreading requires the capacity to empathise. It is only through empathising that one comes to learn what others are thinking and feeling, but these two capacities are not one and the same thing, as in the previous case. Instead, empathy is understood here as a process of simulation that can take (at least) two forms. On the one hand, empathy can be viewed as a low-level form of simulation, which broadly refers to a process of involuntary motor and emotional resonance. On the other hand, empathy can be characterised as a high-level form of simulation, which implies an explicit imaginative re-enactment of the other's psychological state. I shall now briefly characterise each of these, in turn.

Empathy as a low-level form of simulation is usually understood to take place whenever the attended perception of another's action or emotion triggers an involuntary motor or emotional resonance in the perceiving subject. This brings empathy close to the notions of motor mimicry (for instance, yawning in response to someone else's yawning) and emotional contagion (such as when a baby's crying triggers another baby's crying), which inevitably links it to the so-called mirror neurons. These were discovered, in the 1990s, in the brains of macaque monkeys. They were described as a group of neurons that fire both when a monkey executes an action, and when she observes another individual performing a similar action (di Pellegrino et al. 1992; Rizzolatti et al. 1996). Although their existence has not been demonstrated in humans due to the highly invasive nature of the procedures required to investigate them directly, several studies strongly suggest the presence of analogous mirroring systems in our brains, and their involvement in motor and emotional resonance (Fadiga et al. 1995; Carr et al. 2003; Dapretto et al. 2006; Jabbi, Swart, and Keysers 2007; Rizzolatti and Craighero 2004).

Giacomo Rizzolatti *et al.* (1996) originally defended the idea that the function of mirror neurons was to facilitate "the understanding of motor events" by internally representing the observed action. When the mirror neurons discharge, they allow the individual "to recognize the presence of another individual performing an action, to differentiate the observed action from other actions, and to use this information in order to act appropriately" (Rizzolatti et al. 1996, 137). Shortly after, Vittorio Gallese and Alvin Goldman (1998) were the first to link mirror neurons to mindreading. Although they didn't make use of the term 'empathy' in this particular paper, they did postulate that mirroring mechanisms may have a fundamental role to play in the understanding of others' mental states, by enabling the observer to "mimic ... the mental activity of the target," thus constituting "nature's way of getting the observer into the same 'mental shoes' as the target" (Gallese and Goldman 1998, 497-8). Mirroring mechanisms may thus be a part of the high-level simulation heuristics that, according to theorists such as Alvin Goldman, lie at the basis of most of our folk psychological practices (see below). Mirror neurons themselves, as present in macaque monkeys, were postulated to "represent a primitive version, or possibly a precursor in phylogeny, of a simulation heuristic that might underlie mind-reading [in humans]" (Gallese and Goldman 1998, 498).

In later papers, Vittorio Gallese tried to give mirror neurons a role to play in mindreading independently of high-level simulation routines. He defended the idea that "the concept of empathy should be extended in order to accommodate and account for all different aspects of expressive behavior enabling us to establish a meaningful link between others and ourselves" (Gallese 2003, 176-7). This form of empathy is instantiated, at the subpersonal level, by the activity of mirror mechanisms, which "allow us to appreciate, experience, and implicitly and prereflexively understand the emotions and the sensations we take others to experience" (Gallese 2003, 177). Empathy as characterised by Gallese would thus constitute a low-level form of simulation that enables a pre-reflexive and implicit form of mindreading, allowing others' actions and emotions to become meaningful to us.

Mirroring mechanisms have also been understood to have a role to play in enabling the emergence of mindreading capacities. Jennifer Pfeifer and Mirella Dapretto have argued, for instance, that "being able to *feel* what others feel might be a phylogenetic and ontogenetic precursor to more explicit processes of *reasoning through* what others feel" (Pfeifer and Dapretto 2009, 185, their emphasis). This claim is supported by empirical studies that suggest that autistic patients, who typically present mindreading deficiencies, have abnormally functioning mirror mechanisms (Pfeifer and Dapretto 2009, 190; for a review of this research, see: Oberman and Ramachandran 2007).

Tania Singer (2006) has argued that perceiving or imagining someone in an emotional state can trigger a mirroring process that results in a sharing of affective states between the subject and the target. This sharing of affective states allows us to understand the emotions of others in a fundamentally different way from our understanding of their propositional attitudes (which she takes to include only beliefs, desires, and intentions). Singer thus establishes a difference between the sort of understanding gained through mindreading (which she understands specifically as the attribution of propositional attitudes), and the sort of understanding gained through affect sharing or empathy. However, since she conceives of empathy as an ability that enables an understanding of others' emotions, and since these are mental states as well, we can understand her concept as an example of empathy as enabling mindreading, where this notion is understood in the broad sense we have been following. When putting forward this conceptual framework, Singer (2006) is following the perception-action model introduced by Stephanie Preston and Frans de Waal in their famous (2002) paper on empathy:

> A Perception-Action Model of empathy specifically states that *attended perception of the object's state automatically activates the subject's representations of the state, situation, and object, and that activation of these representations automatically primes or generates the associated autonomic and somatic responses, unless inhibited* (Preston and de Waal 2002, 4, their emphasis)

A perception-action mechanism would thus be any mechanism that ensures that when we perceive (and attend to) another individual's emotional or motor state, a largely unconscious representation of this very state is activated in our minds. This representation, in turn, will trigger the same physiological or psychological reactions we would undergo if we were actually in that state, unless this response is inhibited. As I mentioned in the previous section, Preston and de Waal's paper was an attempt to construct a notion of empathy, and of its ultimate and proximate bases, that would subsume all (or most) of the common intuitions with regards to this notion. The resulting characterisation of empathy encompasses a cluster of different abilities that have a role to play both in social cognition and in prosocial/altruistic motivation. Accordingly, not only do Preston and de Waal defend that a perception-action mechanism lies at the basis of motor mimicry and emotional contagion, they also seem to defend the idea that this sort of mechanism could enable the attribution of mental states to others. Indeed, they assert that cognitive empathy, which as we saw in the previous section is roughly equivalent to mindreading, relies on such a mechanism (Preston and de Waal 2002, 4). We can therefore assert that Preston and de Waal also subscribe a notion of empathy (understood as a low-level form of simulation) as enabling mindreading (what they call 'cognitive empathy').

We have seen different ways in which low-level forms of simulation are conceived as somehow enabling or contributing to the appearance of mindreading abilities. These low-level forms of simulation can be triggered by processes that are high-level, such as imagination, but the resulting motor or emotional resonance is in itself low-level, insofar as it's automatic, involuntary, and often even unconscious. Conceptions of empathy as high-level forms of simulation, in turn, present it as a process whereby we consciously and explicitly place ourselves in the others' 'mental shoes,' imaginatively reconstructing their state of mind as a first step to attributing mental states to them. Alvin Goldman, who uses the term 'empathy' to "denote the process of simulation" (Goldman 1992, 29) is one of the major proponents of this conception of empathy:

> Let us now describe the simulation heuristic in more detail. The initial step, of course, is to imagine being 'in the shoes' of the agent ... This means pretending to have the same initial desires, beliefs, or other mental states that the attributor's background information suggests the agent has. The next step is to feed these pretend states into some inferential mechanism, or other cognitive mechanism, and allow that

> mechanism to generate further mental states as outputs by its normal operating procedure ... More precisely, the output state should be viewed as a pretend or surrogate state, since presumably a simulator doesn't feel the very same affect or emotion as a real agent would. Finally, upon noting this output, one ascribes to the agent an occurrence of this output state. (Goldman 1992, 21)

This high-level form of simulation, as described by Goldman, is of course much more complex and cognitively demanding than mere motor or emotional resonance, which seems to be a largely physiological reaction. However, as noted before, low-level forms of simulation may have a role to play in the simulation heuristic. In his 2006 book, Goldman notes that "wherever there is mirroring, the potential for simulation-based mindreading is there, and creatures with the requisite conceptual resources, especially humans, seem to exploit this potential extensively" (Goldman 2006, 140). In any case, Goldman (2006) seems to reserve the term 'empathy' for this high-level process of voluntarily and explicitly putting oneself in the others' mental shoes.

While these considerations suggest that Goldman thinks of empathy as a process that has as its outcome the attribution of mental states to others, and thus, as an ability that enables mindreading, there are some passages where his position is somewhat ambiguous, and he seems to equate empathy and mindreading. For instance, he asserts that "mindreading is an extended form of empathy (where this term's emotive and caring connotation is bracketed)" (Goldman 2006, 4). This quote could be taken as suggesting that Goldman thinks of empathy as mindreading itself. The key here, of course, is in the word 'extended,' which may mean that empathy on its own is not mindreading, and that we need to feed its outputs into some other mechanism, like an inferential process, for empathy to result in mindreading. This interpretation is backed up by the fact that, throughout this book, Goldman constantly uses empathy as a synonym for simulation, and, according to his theory, the simulation itself isn't mindreading, but rather, the process that enables us to attribute mental states to others. In his own words, "empathy, or simulation, is psychologically the most primitive and pervasive method for identifying mental states in others" (Goldman 2006, 296). It thus seems plausible to characterise Goldman's notion of empathy as enabling mindreading.

The notion of empathy as a moral emotion that I will construct in the following chapter will bear a slight resemblance to some of the low-level-simulation conceptions of empathy we have seen in this section. In particular, the notion of emotional resonance will have an important role to play in my upcoming arguments, with the important difference that my notion of empathy will be a type of moral motivation, and not a mechanism for social cognition. Characterisations of empathy as high-level forms of simulation will be left completely aside in the following chapters, as they are (presumably) too cognitively demanding to be of any interest when discussing the minds of nonhuman animals.

4.1.3. Empathy as presupposing mindreading

Having looked at characterisations of empathy that understand it as a synonym for mindreading, and as a mechanism that enables mindreading, I am now going to briefly present one of the most common ways of understanding the relationship between empathy and mindreading. In the three rather different examples we shall see in this section, empathy is a mechanism that gives rise to a special form of interpersonal understanding, and *presupposes*, rather than enables, mindreading. In the characterisations of empathy we shall see in this section, mindreading is a necessary cognitive requirement for empathy to take place, but empathy is no longer reduced to mindreading, as was the case in the definitions we saw in section 4.1.1.

Jean Decety and Philip Jackson (2004) provide a paradigmatic account of empathy as a simulation-based ability that presupposes mindreading. They consider empathy to be what enables us to access a kind of interpersonal understanding that takes the phenomenological form of "a sense of similarity between the feelings one experiences and those expressed by others"

(Decety and Jackson 2004, 71). From a functional point of view, they define empathy as an experience that involves the dynamic interaction of three different components: “affective sharing between the self and the other,” “self-other awareness,” and “mental flexibility to adopt the subjective perspective of the other” (Decety and Jackson 2004, 75). They think of empathy as a process with affective resonance at its core, but one in which there is a clear self-other distinction, and an understanding of the other’s affective state. The idea that their account requires the exercise of one’s mindreading abilities is explicitly backed up by several claims the authors make throughout this paper. They state that “empathy involves ... some minimal recognition and understanding of another’s emotional state” (Decety and Jackson 2004, 73), and also that “empathy is not a simple resonance of affect between the self and other,” but rather “involves an explicit representation of the subjectivity of the other” (Jean Decety and Jackson 2004, 90). It therefore seems clear that empathy as they understand it requires the exercise of our mindreading capacities.

A more ambiguous simulation-based account of empathy as presupposing mindreading can be found in Frédérique de Vignemont and Pierre Jacob’s (2012) discussion of the specific case of empathetic pain. Their characterisation of empathy is composed of five different conditions, according to which, X can be said to empathise with Y’s pain iff:

> i) *Affectivity condition*: X is in some affective state or other *s**;
>
> ii) *Interpersonal similarity condition*: X’s affective state *s** stands in some similarity relation to Y’s affective state *s*;
>
> iii) *Causal path condition*: X’s being in state *s** is caused by Y’s being in state *s*;
>
> iv) *Ascription condition*: X is aware of Y’s being in *s*.
>
> ...
>
> v) *Caring condition*: X must care about Y’s affective life. (de Vignemont and Jacob 2012, 306-7)

Conditions i)-iii) make this notion of empathy a simulation-based one, where the simulation takes a high-level form. For X to empathise with Y, the attended perception of Y's affective state must trigger a similar affective state in X, but the latter will not be initiated by affective resonance, but by an imaginative reenactment of what it would be like to be in the other's mental shoes (de Vignemont and Jacob 2012, 297-9). Condition iv), the ascription condition, is what determines that de Vignemont and Jacob's account of empathy presupposes mindreading as a necessary cognitive ability that must be exercised for empathy to take place. In order to be aware that Y is in a certain affective state *s*, X must exercise her capacity to conceptualise and attribute that state to Y, which means X has to mindread, for affective states are a type of mental state. The authors later substitute condition iv) for the following condition: "iv*) X's being in *s** makes X aware that her being in *s** is caused by Y's being in *s*" (de Vignemont and Jacob 2012, 306). While condition iv*) adds more complexity to the original ascription condition, it still presupposes an exercise of X's mindreading capacities, for becoming aware that one's being in mental state *s** is caused by Y's being in mental state *s* still requires attributing mental state *s* to Y.

De Vignemont and Jacob's condition v), the caring condition, is meant to account for the fact that personal and contextual factors modulate our empathetic responses, so that a negative attitude or an indifference towards Y may preclude X from undergoing empathy. Despite the inclusion of the caring condition among the requirements for empathy, de Vignemont and Jacob are not attempting to answer the question of what moves a person to respond with care or concern towards another person's suffering. They explicitly locate their construct as an *explanans* in the other debate that Batson (2009) associated with the notion of empathy, namely, the debate on how we come to know what others think and feel. However, they distinguish their notion of empathy from "standard" mindreading, and instead characterise it as "*affective* mind reading." In their own words:

> The motivational role of empathetic pain for moral and prosocial behavior (i.e.,

> responding to another's distress to alleviate it) has often been stressed. Our definition of empathetic pain, however, tackles a more fundamental issue, namely, the ability to ascribe pain to other people ... Empathy is a special kind of third-person mind reading, that is, a special affective way of representing and understanding another's affective psychological state (hereafter, *affective* mind reading, by contrast to standard mind reading). (de Vignemont and Jacob 2012, 310)

As we've seen happening with several other notions of empathy, it is hard to elucidate exactly where in our classification de Vignemont and Jacob's would fit. The fact that they themselves consider it as a special type of mindreading suggests that this might be a case of empathy as a synonym for mindreading. I have chosen to present it as an example of empathy presupposing mindreading because their conception seems to require the effective exercise of our mindreading abilities, together with several other conditions, for empathy itself to take place. The resulting outcome is not merely an attribution of an affective state to another (as would be the case if it were simply empathy as mindreading), but one that also involves sharing the affective state of the target and caring for her. This is coherent with their claim that their notion of empathy "involves both affective sharing and affective mind reading" (de Vignemont and Jacob 2012, 296), and also with their claim that it requires "[imagining] the unpleasantness of another's pain" (de Vignemont and Jacob 2012, 312), which presumably cannot be realised without exercising one's mindreading capacities—without first realising that the other is in pain.

If my analysis has been correct, de Vignemont and Jacob's notion would be an example of a conceptualisation of empathy that presupposes mindreading and is simulation-based, as it involves the mental recreation of what it feels like to be in a certain mental state (in this case, pain). Not all accounts of empathy as an ability that presupposes mindreading need necessarily be simulation-based. Shaun Gallagher (2012), for instance, has proposed a notion of empathy that is independent of simulation-related abilities while presupposing mindreading. He states:

> [O]ne can conceive of empathy as being (1) a primary, non-reducible, other-directed feeling of concern or interest that (2) is characterized by a clear distinction between empathizer and the other person, that (3) targets the other's situated experience and (4) consciously ascribes that experience specifically to that other. (Gallagher 2012, 376-7)

Gallagher does not require any isomorphism between the affective states of the subject and the target for empathy to take place—they can both be experiencing affective states that are very different in their phenomenology. Instead, he considers that "the empathizer must be concerned with the target's affective life because of context" (Gallagher 2012, 362), where this concern may take either a positive or a negative form. What is important here is not so much the phenomenal character of the affective states of subject and target, but rather, the presence of a certain narrative competence that enables the empathiser to understand the target's situation (Gallagher 2012, 369). We can empathise with others "only when we can frame their behavior in a narrative that informs us about their history or their situation" (Gallagher 2012, 370). When empathising, we don't take up a secondary affective state. Instead, empathy, according to Gallagher, is a *sui generis* affective state; an "intersubjective affect" that "involves my feeling of being with you with respect to your situated experience" (Gallagher 2012, 375).

One could well understand Gallagher's notion of empathy as presupposing mindreading, for the empathiser's concern must be directed at the other's "situated experience," which would, more often than not, include her mental states. Accordingly, Gallagher claims the following:

> That there has to be some kind of ascription seems right, although it does not have to be an ascription of an affective state (it could be an ascription of a cognitive state, as in the case of empathizing with the intellectual difficulty someone is having). (Gallagher 2012, 376)

Both the ascription of an affective state and the ascription of a cognitive state to another would involve mindreading. This interpretation is also in line with Gallagher's portrayal of the difference between empathy and sympathy, where the former is characterised by a similar intentional structure between the subject's and the target's affective state, in contrast to the latter, where the intentional structure is dissimilar. Thus, while in the case of sympathy, "A feels sad *for* B, who is sad (and perhaps outraged) about an injustice done to B," in the case of empathy, "A feels sad (and/or outrage) *about the injustice done* to B, knowing that B also feels sad (and perhaps outrage) about the injustice done to her" (Gallagher 2012, 360, his emphasis). It's hard to see how the requirement that the empathiser know that the target is also sad could be met without exercising one's mindreading abilities.

At this point, anyone familiar with Gallagher's work might object. His well-known criticism of mainstream social cognition theories and the conception of mindreading that comes with them may be seen as implying that his theory on empathy does not involve taking any stance in the mindreading debate at all, since he seems to reject all form of mindreading—an objection that could also apply to my considerations on Dan Zahavi's work (see section 4.1.1). These two authors strongly criticise any theory that presupposes that mental states are hidden from perception and can only be accessed through inference or simulation, and so stand in opposition to any conception of mindreading that does not allow for us to directly experience others' mental states (see, e.g.: Gallagher and Zahavi 2012, chap. 9). They claim the following:

> The perception of emotion in the movement of others ... does not involve taking a theoretical stance or creating a simulation of some inner state ... This kind of perception-based understanding, therefore, is not a form of mind-reading. In seeing the actions and expressive movements of the other person one already sees their meaning; no inference to a hidden set of mental states (beliefs, desires, etc.) is necessary. (Gallagher and Zahavi 2012, 188)

Their claim that the perception-based understanding of others they advocate “is not a form of mind-reading” could be taken to mean that they reject the appeal to any form of mindreading whatsoever when explaining interpersonal understanding, and so that it does not make sense to consider Gallagher’s notion of empathy as presupposing mindreading and Zahavi’s as a form of mindreading. However, they may be using the term “mind-reading” in a rather narrow sense, to mean only those forms of mindreading that involve an “inference to a hidden set of mental states.” Even if they reject all forms of interpersonal understanding that are theoretically-mediated or rely on indirect inferences instead of direct experience, their theory also seems to require the postulation of some sort of ability that allows us to “access the life of the mind of others in their expressive behaviour and meaningful action” (Gallagher and Zahavi 2012, 191).

While it may be true that I can perceive your embarrassment directly when I see you blushing, there has to be some sort of capacity in me that allows me to interpret it as such. After all, a bee could, perhaps, also perceive the change in colour of your cheeks, but it is doubtful that she would understand that you are embarrassed. This would be the case even if human mindreading were entirely independent of inferential capacities, because there is a more fundamental reason behind the bee’s inability to perceive your embarrassment, namely, that she lacks the adequate conceptual repertoire to perceive your blushing as meaningful. Shannon Spaulding (2015), following Fred Dretske (1969), has made a similar point:

> Simple seeing, also called non-epistemic seeing, is possible for any creature with a functioning visual system. Seeing that, also called epistemic seeing, additionally requires conceptual resources. For example, an insect can see a tennis ball insofar as it can visually discriminate the tennis ball from other objects in the environment. However, the insect cannot see *that* the object is a tennis ball. To see *that* it is tennis ball requires that one have the concept TENNIS BALL, which of course the insect lacks. (Spaulding 2015, 473, her emphasis)

Thus, the bee may be able to visually discriminate the change in colour in your face, as so she may be able to (non-epistemically) see the same as I. However, she cannot see *that* you are embarrassed, because that would require possessing the concept EMBARRASSMENT. When authors such as Gallagher and Zahavi claim that we can see someone's embarrassment in her blushing, they are most likely referring to *epistemic* seeing, instead of non-epistemic seeing, because only the former can provide "an alternative explanation [to standard mindreading accounts] of how we *understand* others' behavior" (Spaulding 2015, 473, her emphasis). Gallagher and Zahavi could be hinting at this when they assert that "we need concepts in order to extract and comprehend the informational richness of what is already given, already present to us" (Gallagher and Zahavi 2012, 172). Their theories may thus be compatible with a broader notion of mindreading abilities, understood as whatever allows us to comprehend that those with whom we interact are beings with mental lives.

Whether my interpretation of Gallagher and Zahavi's work is correct or not is, fortunately, irrelevant for present purposes, for my interest lies precisely in arguing that being a moral subject does not require mindreading capacities. In the following chapter, I will construct a notion of empathy as a moral emotion that will be entirely independent of any capacity to interpret others as mental beings. It is therefore not necessary for me to take any stand on how mindreading capacities are implemented in human beings. Additionally, all the notions of empathy we've seen in this section can be left entirely aside, since they presuppose mindreading, and thus go directly against my interests. Even if my interpretation of Gallagher's work were wrong, his characterisation of empathy is too intellectually demanding to be of any interest when discussing the motivations behind the behaviour of nonhuman animals, as it seems safe to assume that the latter lack the sort of narrative competence that Gallagher requires for empathic understanding of others to occur.

4.2. EMPATHY AS A MOTIVATION

Debates on moral psychology attempt, amongst other things, to elucidate what moves us to engage in sensitive care in response to another's suffering. Empathy is often used here as an *explanans*. What is emphasised in this case is not the role that empathy plays in interpersonal understanding, but rather, the behavioural response that empathy can trigger. This understanding of empathy as a motivation for caring behaviour has its origins in the term 'sympathy' as it appears in the works of David Hume and Adam Smith. According to Hume, sympathy is what allows us to "feel ... the pains and pleasures of others," which, in turn, "interests us in the fortunes of others," and can be a motivation for helping them or caring about them (Hume [1739] 1992, 2:169). Interestingly, Hume considers that sympathy is not restricted to the human species. He writes:

> 'Tis evident, that *sympathy*, or the communication of passions, takes place among animals, no less than among men. Fear, anger, courage and other affections are frequently communicated from one animal to another, without their knowledge of that cause, which produc'd the original passion. Grief likewise is receiv'd by sympathy; and produces almost all the same consequences, and excites the same emotions as in our species. The howlings and lamentations of a dog produce a sensible concern in his fellows. (Hume [1739] 1992, 2:180)

In the case of Adam Smith, he defines sympathy as "our fellow-feeling with any passion whatever" (Smith [1759] 1982, sec. I.i.i.5). This "fellow-feeling" may arise by way of the perception of an emotion in another individual, or by exercising our imagination and putting ourselves in the other's place. Sympathy, in Smith's view, constitutes one of the principles in man’s nature that “interest him in the fortunes of others, and render their happiness necessary to him, though he derives nothing from it except the pleasure of seeing it” (Smith [1759] 1982, sec. I.i.i.1). While both Hume and Smith use the term ‘sympathy’ instead of ‘empathy’ (recall that ‘empathy’ was not introduced in the English language until the beginning of the 20th century), the definitions they subscribe bring their notions closer to the contemporary understanding of empathy, which, in contrast to sympathy, is associated with affective resonance or state-matching. In Hume and Smith, this process of resonating with another’s emotion is not a mechanism for understanding others, but a source of concern or care. Contemporary notions of empathy as a moral motivation thus have their intellectual roots in these two authors.

Within current debates on what moves us to help or care for others, the term ‘empathy’ is also used rather ambiguously, although there is a tendency towards characterising it as a multi-level process with affective resonance at its core. If we look at the relationship between empathy and mindreading in this debate, the conceptual landscape appears here largely dominated by a dichotomy: either (1) one possesses (fully developed) mindreading skills and can thus engage in empathically-motivated moral behaviour, or (2) one lacks (fully developed) mindreading skills, and can thus only be moved to act by lower forms of empathy that do not result in moral behaviour. I will now offer some examples of how this dichotomy is presented in the literature. Note that the term ‘moral behaviour’ will be used throughout this section in a rather broad sense that encompasses altruistic behaviour, prosocial behaviour, and more generally, any caring response to another’s suffering. In the following chapter, I will narrow this definition down, but for present purposes, this broader characterisation suffices.

When looking at empathy as a motivation, it is common to find authors who understand it as a multi-level ability, with involuntary physiological phenomena at the lowest levels, and highly complex cognitive processes at the highest ones. The high-level forms of empathy will thus be considered much more cognitively demanding than the lower ones, and moral behaviour will be associated with those higher-level forms. Martin Hoffman (1990), for instance, defines empathy as "an affective response more appropriate to someone else's situation than to one's own situation" (Hoffman 1990, 152). This is, of course, quite a vague characterisation that can apply to many different phenomena, and indeed, Hoffman distinguishes five "modes" of empathy, which correspond to the different ways in which one's affect can be "more appropriate" to the other's situation than one's own. Three of these modes are automatic and involuntary ("primary circular reaction," "mimicry," and "conditioning and direct association") and two are higher-order cognitive ("language-mediated association" and "putting self in other's place") (Hoffman 1990, 154). Hoffman also distinguishes four different ways in which we can understand others ("self-other fusion," "other represented as physically distinct from self," "other represented as having internal states independent of one's own," and "other represented as having ... his/her own history and identity") (Hoffman 1990, 154). These ways we can understand others, in turn, correspond to four different developmental levels of empathy:

1. The first developmental level refers to what Hoffman terms "global empathy," which corresponds to the most basic form of empathic distress—so basic that it can be found in infants. Global empathy refers to a phenomenon of automatic and involuntary emotional contagion in which there is a "self-other fusion," so that "distress cues coming directly from the dimly perceived other person are confounded with unpleasant feelings empathically aroused in the self" (Hoffman 1990, 153-4). At this level there is no mindreading and the resulting response can never be moral, since it will be entirely self-centred. This is why infants in the presence of someone in distress will tend to "act as though what happened to the other person happened to themselves" (Hoffman 1990, 155).
2. The second level, "'egocentric' empathy," emerges when children start to understand others as beings who are "physically distinct from themselves." At this stage, however, there is still no mindreading, so that the other's mental states "remain unknown to the child and continue to be confused with the child's own internal states" (Hoffman 1990,

155). Accordingly, young children will often offer those in distress whatever would comfort themselves, such as their own toy.

3. The third level is termed "empathy for another's feelings" and comes with the development of mindreading and role-taking abilities. This allows children to become "more responsive to cues about what the other is actually feeling," which allows for the rise of help that is directed at and tailored to the other's specific needs. One also starts to empathise with more complex emotions, instead of merely distress (Hoffman 1990, 155).
4. The fourth, and final, level is termed "empathy for another's life condition," and emerges when one starts to understand others as beings with a specific history and identity. This allows, not only for our empathic response to be intensified when we realise that someone's distress is "not transitory but chronic," but also, for the development of sophisticated moral judgement capacities. For example, it is only when we have reached this stage that we can adhere to "political ideologies centered around alleviation of the plight of disadvantaged groups" (Hoffman 1990, 155-6).

In Hoffman's framework, therefore, the development of caring responses that are tailored to the other's specific needs, and of sophisticated moral judgement capacities, is closely linked to the development of mindreading capacities. Empathy here is no longer a mechanism for interpersonal understanding. Instead, its importance comes by way of its role in providing "the affective and motivational base for moral development and just behavior," thus constituting "a major cohesive force or glue in society" (Hoffman 1990, 151).

Frans de Waal also champions a conception of empathy as a multi-level phenomenon that has a big role to play in "the regulation of social interactions, coordinated activity, and cooperation toward shared goals" (de Waal 2008, 282). He defines empathy as "the capacity to (a) be affected by and share the emotional state of another, (b) assess the reasons for the other's state, and (c) identify with the other, adopting his or her perspective" (de Waal 2008, 281). Thus defined, empathy appears as quite cognitively demanding, and would surely require mindreading capacities. However, de Waal considers that empathy would also take place if just condition (a) were met. This is because he follows what he calls the "Russian Doll model" (see figure 3). At the core of all empathic processes lies the perception-action mechanism that was introduced in Preston and de Waal (2002). The outer layers of the Russian Doll correspond to higher-level abilities such as mindreading and perspective-taking skills. This results in three different levels of empathy, all of which will be underpinned by a perception-action mechanism that will provide the adequate motivation for the corresponding behavioural outcomes (de Waal 2008, 287). The Russian Doll model also applies to the different forms of imitation, which de Waal considers to be underpinned by a perception-action mechanism as well.

What de Waal considers to be "[t]he lowest common denominator of all empathic processes" is emotional contagion, which occurs when "one party is affected by another's emotional or arousal state" (de Waal 2008, 282). Emotional contagion is the automatic adoption

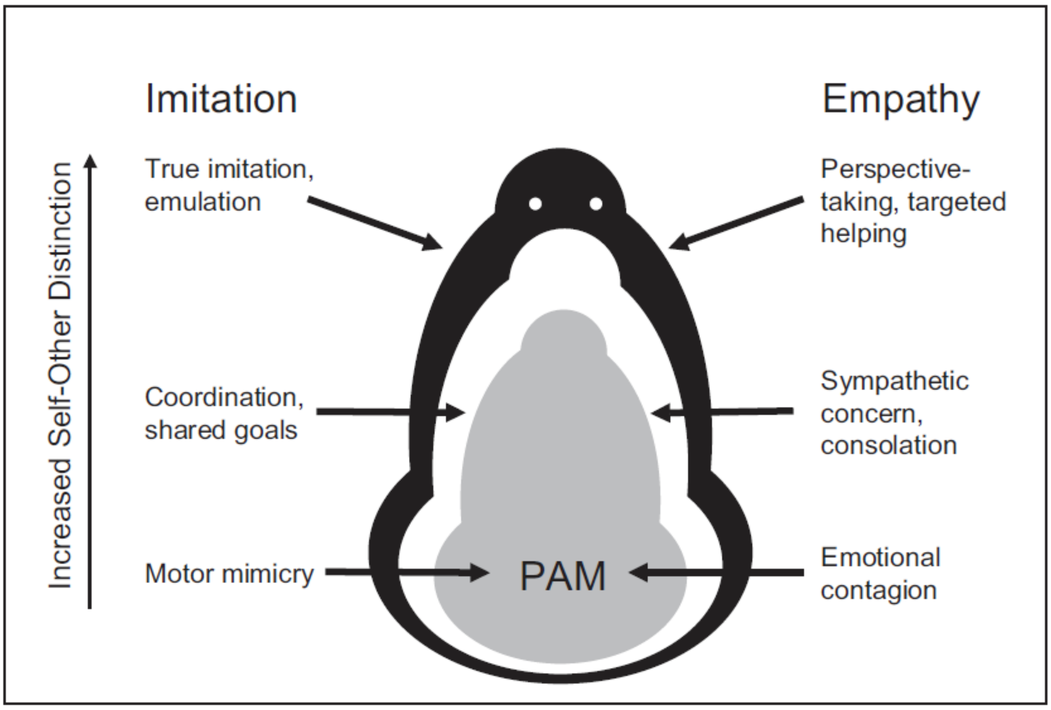
Imitation
Empathy
Increased Self-Other Distinction
True imitation, emulation
Perspective-taking, targeted helping
Coordination, shared goals
Sympathetic concern, consolation
Motor mimicry
PAM
Emotional contagion

Figure 3: Frans de Waal's Russian Doll model. Source: de Waal (2008, 288).

of another's emotional state, and does not require any mindreading skills or understanding of what triggered the original reaction. Emotional contagion is likely what causes a baby's crying in response to another baby's crying (Sagi and Hoffman 1976), or a flock of birds' simultaneously taking off when one of them is startled and flies away. Emotional contagion may collapse into personal distress, in which case the resulting emotion will be entirely self-centred, and will make "the affected party selfishly seek to alleviate its own distress, which mimics that of the object" (de Waal 2008, 283). This is all the more likely if the affected party lacks the ability to distinguish self from others, and also if she lacks mindreading skills.

The next level is constituted by what he terms "cognitive empathy," which "occurs when emotional contagion is combined with appraisal of the other's situation and attempts to understand the cause of the other's emotions" (de Waal 2008, 283). This results in a feeling of "sympathetic concern" that is directed at the target's emotional state, and may give rise to an attempt to act upon the situation in order to ameliorate it. Note that mindreading skills must be present for sympathetic concern to occur. In contrast to personal distress, sympathetic concern "relies on a separation between internally and externally generated emotions" (de Waal 2008, 284). De Waal states that sympathetic concern is likely at the basis of the examples of consolation behaviour found in nonhuman animals (see section 2.1.2.3).

The third and final level refers to "empathic perspective-taking." Perspective-taking is a cognitive endeavour that implies adopting another's point of view, and requires exercising one's imagination and mindreading skills. Perspective-taking takes on the adjective 'empathic' when it also incorporates "emotional engagement," so that the term 'empathic perspective-taking' refers to "the capacity to take another's perspective—e.g., understanding another's specific situation and needs separate from one's own—combined with vicarious emotional arousal" (de Waal 2008, 285). This high-level form of empathy can trigger behavioural responses that are tailored to the specific situation of the individual in need—what de Waal terms "targeted helping," which he defines as "help and care based on a cognitive appreciation of the other's specific need or situation" (de Waal 2008, 285). According to de Waal, empathic perspective-taking, which he takes to be present in great apes, is more likely to occur when individuals possess a heightened sense of self. In his view, those animals capable of passing the mirror self-recognition test devised by Gordon Gallup, Jr. (1982) will likely posses this form of empathy.

Just like Martin Hoffman, de Waal links the appearance of sophisticated caring behaviour to the presence of higher-order capacities such as mindreading skills. In contrast to Hoffman, however, de Waal explicitly considers the possibility that some nonhuman animals (most notably, great apes) may sometimes behave on the basis of these higher-level forms of empathy. He discards the possibility that monkeys may possess sympathetic concern or empathic perspective-taking, precisely due to the fact that studies tend to suggest that great apes, but not monkeys, possess mindreading skills (de Waal 2008, 285). While he considers that great apes may possess the higher forms of empathy, he stays clear of the use of the term 'moral' in this paper. Elsewhere, he argues that we can find the "building blocks" of human morality in some nonhuman primates, where these building blocks include "the capacity for empathy, a tendency for reciprocity, a sense of fairness, and the ability to harmonize relationships" (de Waal 2006, 168). Human morality would consist of two further levels, constituted by mechanisms of social pressure and by our moral judgement skills. Nonhuman primates are considered by de Waal to lack these two higher levels. He avoids, however, taking an explicit stand on whether their behaviour qualifies as moral or not:

> To neglect the common ground with other primates, and to deny the evolutionary roots of human morality, would be like arriving at the top of a tower to declare that the rest of the building is irrelevant, that the precious concept of "tower" ought to be reserved for its summit ... Are animals moral? Let us simply conclude that they occupy several floors of the tower of morality. Rejection of even this modest proposal can only result in an impoverished view of the structure as a whole. (de Waal 2006, 181)

Beth Dixon (2008) argues that de Waal has not succeeded in building a case for morality, or even proto-morality, in animals, and, in doing so, she offers a characterisation of empathy as a moral motivation. She writes:

> De Waal fails to specify in the very definition of cognitive empathy what is morally relevant about this emotional state. Empathy may involve appraising another's situation precisely, but what matters to empathy understood as a moral concept is that the subject perceives what is morally salient about another's situation. Even this additional requirement doesn't quite capture what needs to be added to cognitive empathy to make it a morally significant concept, and that is its relation to sympathy or compassion. These states are genuine moral emotions in the case where they motivate a subject to help or to alleviate need, distress, or suffering when this is judged to be "serious" or undeserved. The moral significance of the emotional states of sympathy and compassion is explained by the presence of evaluative judgments as well as the motivations to act on these evaluations or appraisals. (Dixon 2008, 140)

We've seen that, in de Waal's framework, it is not until the appraisal that comes with cognitive empathy is added to emotional contagion, that we can start to speak of moral (or proto-moral) forms of empathy. This appraisal would include an appreciation of the context in which the other's distress is taking place, and an understanding of what caused it (de Waal 2008, 283). According to Dixon, however, emotional contagion combined with this sort of appraisal would not be enough to make it a moral reaction. For empathy to count as a moral emotion, two more things are needed: first, the presence of an explicit evaluative judgement that entails that the subject considers the target's distress to be "'serious' or undeserved," and second, the motivation to act in accordance with both this evaluation and the contextual appraisal. Thus, Dixon not only requires the presence of mindreading capacities, but also the ability to engage in moral judgements and to be motivated to act by them. In her view, de Waal does not succeed in capturing the moral character of empathy. Moreover, he is wrong to think that animals could be part of the (proto-)moral realm because the conditions required to turn empathy into a moral emotion are too intellectually demanding to be fulfilled by any other species than our own.

Some conceptions of empathy as a moral motivation are not simulation-based, but still presuppose the presence of mindreading capacities. Daniel Batson, for instance, has written:

> Feeling for another person who is suffering ... is the form of empathy most often invoked to explain what leads one person to respond with sensitive care to the suffering of another ... To feel for another, one must think one knows the other's internal state ... because feeling *for* is based on a perception of the other's welfare (Batson 2009, 9–10)

Batson is referring here to a form of empathy that he defines as "an other-oriented emotional response elicited by and congruent with the perceived welfare of someone else" (Batson 2009, 8). The term 'other-oriented' means that the target is the intentional object of the emotion, and 'congruent' refers to the fact that the valence of the subject's emotion (i.e. whether it is positive or negative) has to coincide with the valence of the target's perceived welfare. This form of empathy is not a result of affective resonance, and so a matching of affective states between subject and target is not required. Feeling *sad* for a friend who is *scared* would be an example of empathy thus understood. It is a form of concern that emerges as a result of a *believing that* the target is suffering. While this kind of empathy can be a motivation to engage in caring or helping behaviour, it is dependent upon one's mindreading capacities to emerge in the first place. Without believing that the target is distressed or in pain, one cannot be concerned for her welfare.

Kristin Andrews and Lori Gruen (2014) have given an account of empathy as a moral motivation that aims to de-intellectualise this capacity and allow for nonhuman apes to possess it. While they do construct a notion of empathy that is less intellectually demanding than others (for instance, Dixon's), they still rely on the presence of mindreading capacities in the empathising subject. As is common in the literature, they rule out the possibility that bare emotional contagion—which they view as the most basic form of empathy—may have anything to do with morality. They define it as "a kind of mimicry of the individual(s) in one's immediate environment" that "does not require any developed cognitive capacities" (Andrews and Gruen 2014, 195). They argue that there can be no role for this kind of capacity in morality, because "the empathizer isn't distinguishing his or her own feelings or mental states more generally from those of another" (Andrews and Gruen 2014, 195). However, they point out that it is not necessary, in order to account for empathy-based moral behaviour, to construct a highly intellectualistic and detached notion of empathy that would require "taking the perspective of another in order to understand what that other is experiencing and making decisions about what to do in light of what the other is experiencing" (Andrews and Gruen 2014, 195). The type of empathy they advocate for—which, following Gruen (2015), they call "entangled empathy"—stands somewhere in between these two capacities:

> [E]ntangled empathy ... involves being able to understand and respond to another's needs, interests, desires, vulnerabilities, and perspectives not as if they are or should be the same as one's own. It involves a reaction to another's experience and a judgment to act in response. This latter form of empathy has a clearer connection to ethics (Andrews and Gruen 2014, 196)

This definition quite strongly suggests that entangled empathy requires the presence of mindreading capacities. Without them, we could never understand that others have "desires" and "perspectives." Moreover, understanding someone as possessor of specific "needs," "interests," and "vulnerabilities," requires, at the very least, understanding that she possesses sentience; that, for her, some things are pleasant and others unpleasant, and so, that she prefers undergoing some experiences and not others. This sort of understanding cannot be reached without representing the target as a being with a mental life who can feel pain, suffer, and so on. While Andrews and Gruen de-emphasise the importance of the attribution of beliefs to others, they explicitly state that entangled empathy requires considering "others' physical or social situation, their capability, their emotions, and their differing goals," as well as "seeing others as somewhat different from oneself, and from one another" (Andrews and Gruen 2014, 202). Mindreading abilities would undoubtedly have to be present for an individual to develop such a nuanced understanding of others.

As we've seen by looking at these different definitions of empathy, there seems to be a widespread consensus in the literature with respect to the idea that mindreading is required for empathy to count as a moral motivation, which means that, without mindreading, empathy cannot generate behavioural responses that can be characterised as moral(/prosocial/altruistic). This general agreement prevails even if there is a divergence in the characterisations of empathy as a moral motivation, so that we can even find examples of apparently empathy-based prosocial behaviour in chimpanzees being presented as evidence of mindreading capacities in this species (O'Connell 1995). My aim in the next chapter will be to directly confront this consensus and show how we can conceive of a form of empathy that will not require mindreading capacities, yet still count as a moral emotion. But before moving on, it is important to present the way in which the empirical evidence for animal morality has often been reinterpreted to accommodate this theoretical commitment, namely, that moral forms of empathy require mindreading.

4.3. EMPATHY IN ANIMALS: MORALITY OR AVERSIVE AROUSAL?

Let us begin this section by recalling a study performed by Inbal Bartal, Jean Decety, and Peggy

Mason (2011), which we already saw in chapter 2. In this experiment, a rat was placed in an enclosure where a fellow rat was trapped inside a restrainer. The experimenters found that, once the free rat learnt how the restrainer could be opened, she would do so at a short latency and in a reliable manner whenever there was a fellow rat trapped in it, but not in those cases in which the restrainer was empty or there was a toy rat inside. The experimenters ruled out the hypothesis that the rat's performance was due to an expectation of social contact with the trapped cagemate by setting the apparatus so that the latter would be liberated into a separate compartment, which did not result in a change in the free rat's behaviour. When a second restrainer full of chocolate chips (which rats have a strong preference for) was placed in the compartment, the free rat preferred to release her companion and share the food with her rather than eat it on her own first. The experimental subjects' restrainer-opening behaviour persisted for more than a month, which strongly undermines the possibility that they were acting out of curiosity. Moreover, the fact that opening the restrainer was not easy and yet it occurred at a short latency, indicates that their helping behaviour was not accidental, but intentional.

According to the authors of this study, it is likely that the experimental subjects were behaving on the basis of "the rodent homolog of empathy" (Bartal, Decety, and Mason 2011, 1430). The subjects displayed what seems like a clear example of altruistic behaviour, for they chose to help their companion even when they had nothing to gain, or even something to lose. Does this mean that they were behaving on the basis of a *moral* form of empathy? Using an analogous experimental design, Nowbahari *et al.* (2009) had previously found that ants will reliably free non-anesthetised nest-mates who are trapped in a snare. Would we want to say that ants can behave morally too? There seems to be an intuitive difference between the performance of these two species, even though, on the surface, their behaviour appears equitable. To determine which of these two animals, if any, actually behaved morally when rescuing their conspecific, we would have to look at the underlying mechanisms. It would be unjustified to speak of morality basing our claims solely on the external description of these individuals' behaviour. As we shall see in the following chapter, behaviour can only be labelled 'moral' if it has been triggered by a motivation of the appropriate sort.

Following Marco Vasconcelos *et al.* (2012), we can distinguish between behaviour that is *functionally* intended to benefit another and behaviour that is *psychologically* intended to benefit another. Both the ants' and the rats' behaviour qualify as the former, that is, they are both "designed by natural selection to benefit another" (Vasconcelos et al. 2012, 911). However, for their behaviour to be psychologically (and not just functionally) intended to benefit another, according to Vasconcelos *et al.*, it would have to have been "goal-directed towards acting on the actor's *internal representation of the receiver's wellbeing*" (Vasconcelos et al. 2012, 911, my emphasis). A way of preserving the morality of the rats' behaviour, while precluding the ants' from qualifying as moral, would be to state that only the former was psychologically intended to helping the conspecific in need. This ties in with the position of the authors we saw in the previous section, who would presumably consider that the rats' behaviour could only have been motivated by a *moral* form of empathy if rats were capable of attributing mental states to the trapped cagemate—a necessary requirement if we want to describe their performance as intentionally aimed at reducing their companion's distress. And yet, it seems *prima facie* unlikely that rats possess such an intellectually demanding capacity. Following this framework, this would mean that their behaviour cannot be moral. The question then becomes: how do we reinterpret the data of this and other studies to accommodate this intellectual commitment—that moral behaviour requires mindreading—? And how do we explain the intuitive difference between the rats' performance in Bartal *et al.* (2011) and the ants' in Nowbahari *et al.* (2009)?

In 1962, George Rice and Priscilla Gainer found that rats faced with a conspecific suspended from a harness and displaying signs of distress would reliably choose to press a bar that lowered their conspecific to the floor. The bar was not pressed when the harness contained a styrofoam block instead of a rat. According to Rice and Gainer, these results suggested that rats can engage in a behaviour that is "homologous to altruism" (Rice and Gainer 1962, 124). In an article written in 1963, J.J. Lavery and P.J. Foley argued that this conclusion did not follow from the experimental results. In a similar argument to the one made later by Vasconcelos *et al.* (2012), Lavery and Foley defended that the rats' behaviour cannot be termed 'altruistic' "unless it can be demonstrated that the bar is pressed specifically to relieve the distress of the other animal" (Lavery and Foley 1963, 172). Unless the helping rat understands that the other animal is distressed (i.e. unless she possesses mindreading skills), and aims specifically at alleviating her suffering, her behaviour cannot be considered altruistic.

Lavery and Foley set out to test an alternative hypothesis, namely, that the distressed rat's cries were simply a source of auditory stimulation that resulted in an arousal, which, in turn, led to increased behavioural activity in the experimental subjects. To test this hypothesis, they subjected several rats to recordings of squeals from conspecifics, and to white noise. Both stimuli could be stopped by pressing a bar. The experimenters found that the number of times the bar was pressed over trials was reliably higher during white noise than during squeals, which they interpreted as consistent with their hypothesis. These experimental results have been interpreted elsewhere as evidence that rats experience both white noise and distress cries from conspecifics as aversive stimuli, which would mean that, by pressing the bar, in the original Rice and Gainer experiment, they were simply aiming at alleviating their own distress, and not their conspecific's. Marc Hauser, for instance, has written that "this result shows that rats will press a bar to turn off a variety of unpleasant stimuli, including rat squeals" (Hauser 2001, 273), which points to the idea that, in Rice and Gainer (1962), the rats were aiming to "reduce the unpleasantness of another's noises (a *selfish* act)," rather than to "reduce another's suffering (an *altruistic* act)" (Hauser 2001, 273, his emphasis).

This *aversive arousal hypothesis* has been widely used to account for experimental results that would otherwise point to the presence of empathy-based moral behaviours in nonhuman animals. The reader may recall that one of the first experiments to deliver such suggestive results was Russell Church's (1959), where rats temporarily refrained from pressing a lever to obtain food when this resulted in an electric shock being delivered to a conspecific. Hauser also reinterpreted these results to fit the aversive arousal hypothesis:

> These experiments suggest that rats are willing to eat less if, by inhibiting particular responses, they benefit another rat. This looks like altruism based on sympathy, perhaps even empathy. Several questions remain, however. Perhaps rats stop pressing the lever because the other rat's screaming and wriggling are annoying, physically unpleasant. ... All animals will find ways to convert an unpleasant situation into something pleasant or at least less unpleasant. (Hauser 2001, 272)

Joseph Lucke and Daniel Batson (1980) also reinterpreted Church (1959) and Rice and Gainer's (1962) experimental results along the lines of the aversive arousal hypothesis. They argued that, before claiming that the rats' behaviour is altruistic, it is important to rule out three alternative explanations: (a) the companion's squeals are an unconditioned aversive stimulus, as would be a very loud noise, (b) the companion's squeals are a conditioned aversive stimulus, resulting from prior personal experience with electric shocks, and (c) the subjects are anticipating an imminent shock to themselves, which would have caused them, in Church (1959), to freeze and thus refrain from pressing the lever (Lucke and Batson 1980, 215-7).

To explore these alternative explanations, Lucke and Batson conducted an experiment in which rats were presented with a fellow rat being shocked as a result of the former pressing a bar to obtain food. Some of the subjects had previous experience with shocks (either on their own, or together with their companion), and some had none. The experimenters found that, when neither the subject nor the target had been previously shocked, or when only the subject had previous shock experience, there was a much lower bar pressing suppression than when the subjects had been conditioned to expect an electric shock upon seeing their companion shocked. Lucke and Batson concluded that this experiment "provided no clear evidence for either innate or acquired social concern," and instead suggested "that when laboratory rats help a distressed companion, it is not because of altruistic concern for the other; their concern is for themselves" (Lucke and Batson 1980, 214).

The famous experiment performed by Jules Masserman, Stanley Wechkin, and William Terris (1964), which suggested that "[a] majority of rhesus monkeys will consistently suffer hunger rather than secure food at the expense of electroshock to a conspecific" (Masserman, Wechkin, and Terris 1964, 585), has also been reinterpreted along the lines of the aversive arousal hypothesis. With respect to this experiment, Hauser has written:

> What is most remarkable about these experiments is the observation that some rhesus monkeys refrained from eating in order to avoid injuring another individual. Perhaps the actors empathized, feeling what it would be like to be shocked, what it would be like to be the other monkey in pain. Alternatively, perhaps seeing someone shocked is unpleasant, and rhesus will do whatever they can to avoid unpleasant conditions. Although this has the superficial appearance of an empathic or sympathetic response, it may actually be selfish. (Hauser 2001, 276)

In this fragment, we can also see that Hauser thinks of the kind of empathy that can give rise to altruistic behaviour as perspective-taking ("feeling ... what it would be like to be the other monkey in pain"), and thus, as involving mindreading. The only alternative explanation that Hauser devises for these experimental results is aversive arousal, which would be a "selfish" reaction. Frans de Waal has also suggested the aversive arousal hypothesis as an alternative explanation for Masserman *et al.*'s (1964) results:

> Perhaps the most compelling evidence for emotional contagion came from Wechkin, et al. (1964) and Masserman et al. (1964), who found that monkeys refuse to pull a chain that delivers food to them if doing so delivers an electric shock to and triggers pain reactions in a companion. Whether their sacrifice reflects concern for the other ... remains unclear, however, as it might also be explained as avoidance of aversive vicarious arousal (de Waal 2008, 283)

The aversive arousal hypothesis allows scholars to offer a *non-moral* explanation of the performance of animals in these experiments. Vasconcelos *et al.* (2012) write that "[a]cting to make ... another subject's stress signals cease can be psychologically selfish and does not require empathic interpretations," (Vasconcelos et al. 2012, 911) where empathy is understood as "as a proximate mechanism [that] requires that the agent [rely] on a cognitive or emotional representation of someone else's internal state" (Vasconcelos et al. 2012, 910). In other words, when behaviour has been triggered by aversive arousal, it will be non-moral because the motivation behind it will be simply to eliminate a stimulus that the subject herself experiences as unpleasant. Helping or caring behaviour that is triggered by aversive arousal aims solely at eliminating the subject's own distress by way of the extinction of its cause. The motivation is, therefore, entirely selfish; the concern is only for oneself.

Not only is this interpretation a non-moral explanation of the evidence, it is also a *deflationary* account of the animals' behaviour. It is deflationary because aversive arousal is much less intellectually demanding than moral forms of empathy. It can be triggered by an entirely physiological response, such as emotional contagion, but does not require any understanding of the external cause of the unpleasantness that the individual is feeling. While aversive arousal can have affective resonance at its core, and thus, under some accounts, can count as an empathic reaction, it *doesn't require mindreading skills*.[1] The aroused subject can be a mere behaviour reader, in Povinelli *et al.*'s sense (see section 3.2), and simply experience the

[1] Note that aversive arousal can also occur in the presence of mindreading skills. There is an old and ongoing debate in moral psychology about whether all apparently moral behaviour in humans is ultimately aimed solely at alleviating our own distress when seeing others' suffering, which would render it selfish (see Doris and Stich 2014, sec. 5.5, for an overview of this debate). For the purposes of this dissertation, I will ignore this problem, since my initial presuppositions include the assumption that moral behaviour does indeed exist. While I will not deal with this problem directly, some of the considerations made in the following chapter could be interpreted as indirect contributions to this debate.

superficial cues that come with the other's distress as "bad music that you try to turn off" (Nichols 2001, 429). The dichotomy thus becomes: either (1) the animals in these studies possess mindreading skills and their motivation to help the other could, perhaps, be a moral form of empathy, or (2) they lack mindreading skills, and so, the most likely explanation for their performance is that they experienced the other's distress cues as an aversive stimulus that they wanted to eliminate, which renders their behaviour selfish and non-moral.

Going back to where we started at the beginning of this section, we can now see how the aversive arousal hypothesis would allow us to preserve the difference between the rats' performance in Bartal *et al.* (2011) and the ants' behaviour in Nowbahari *et al.* (2009) without having to attribute highly intellectually demanding capacities to rats. While the ants' behaviour was probably triggered by a "chemical call for help" or some sort of "eliciting stimulus" that was automatically released by the entrapped conspecific (Nowbahari et al. 2009, 3), the rats' behaviour probably had some sort of affective component, such as emotional contagion, as its motivation. The distress of the trapped cagemate perhaps had a contagious effect on the free rat, who seeked to relieve herself of her own distress by releasing her companion. This would render the rat's behaviour somewhat more intellectually demanding than the ants,' without having to postulate, as Bartal *et al.* did, an "ability to understand and actively respond to the affective state of a conspecific" (Bartal, Decety, and Mason 2011, 1430).

While the aversive arousal hypothesis has a certain elegance to it, and would certainly help explain away much of the evidence we saw in chapter 2—thus contributing to preserve the uniqueness of human morality—, I shall argue against it in the following chapter. My strategy won't be to come up with an alternative explanation of the data. Instead, I shall attack the way the aversive arousal hypothesis has usually been interpreted. The common dichotomy between aversive arousal and morality is, I shall argue, a false one. Provided certain conditions are met, aversive arousal (or rather, something very close to it) can actually constitute a moral motivation. If my upcoming arguments are right, this will open up a way in which individuals with no mindreading skills whatsoever can nevertheless be moral.

4.4. SUMMARY

My aim in this chapter has been twofold. First, I have mapped out the different ways in which the relationship between empathy and mindreading can be understood. We have seen that there are two divergent trends in the definitional debates on empathy. There are, on the one hand, scholars who follow in the tradition that stems from Theodor Lipps' work, and who understand empathy as a mechanism involved in interpersonal understanding. On the other hand, many scholars consider empathy to be a motivation that can result in helping or caring behaviour directed at others. These second scholars have an understanding of empathy that stems from David Hume and Adam Smith's notion of sympathy.

We have seen that the theoretical relation between empathy and mindreading is as diverse as the many uses of the term 'empathy.' Within the research strand that deals with social cognition, there are at least three different ways in which empathy is thought to relate to mindreading. Empathy can be understood as (1) a synonym for mindreading, (2) a mechanism that enables mindreading, and (3) a complex psychological process that requires or presupposes mindreading. Within the research strand that deals with moral psychology, there is a divergent understanding of empathy, but a convergent understanding of its relation to mindreading and morality. The consensus seems to be that, if empathy is to count as a moral motivation at all, mindreading skills must be present. This is a claim that I intend to argue against in the following chapter. My position will be that there can be minimal forms of empathy that do not require mindreading capacities and yet still qualify as a moral motivation.

My second aim in this chapter has been to show how the evidence that points to empathy-based moral behaviour in animals has often been reinterpreted to accommodate the commitment to a necessary connection between morality and mindreading. It is commonly understood that, for an animal's helping/caring behaviour to be considered moral, it has to be specifically and psychologically intended at alleviating the suffering of the other. If an animal lacks mindreading skills, she cannot understand that the target is suffering, and so she cannot intend to modify the target's state. The most common way to reinterpret the empirical evidence to fit this claim is by reference to the aversive arousal hypothesis, according to which, those animals who appear to help others out of a concern for their welfare, are actually aiming at improving their own welfare, by eliminating an aversive stimulus that is jeopardising their own wellbeing. What appears to be a moral reaction is actually selfish. This is also a claim that I will argue against in the following chapter, where I will defend the idea that aversive arousal is not necessarily incompatible with morality.

5. MINIMAL MORAL EMPATHY

This chapter contains a defence of the core idea of this dissertation, namely, the idea that morality does not require mindreading capacities. I will defend this thesis by focusing on the specific emotion of empathy. I will show that a minimal form of empathy that does not require the presence of mindreading capacities, and yet, qualifies as a moral emotion is theoretically conceivable and empirically plausible. If my reasoning in the following pages is correct, then we will be able to conclude there is at least one type of moral motivation that lies within the reach of non-mindreading animals.

The arguments in this chapter are located at the intersection of three contemporary debates, which deal, in turn, with animal morality, animal mindreading, and the proper use of the term 'empathy.' In the previous three chapters, I have set the stage for the present one by giving an overview of these three debates, and in doing so, introducing the conceptual tools I will use, and the arguments that I will try to refute. Let us now briefly recall the most important points I made along the way.

Chapter 2 focused on the animal morality debate. I showed that, despite the growing amount of empirical evidence that points to the presence of moral behaviour in animals, most scientists and philosophers remain altogether sceptical with respect to the possibility that any nonhuman creature may be moral. We saw that this tension was at least partly due to the compelling nature of the *amalgamation fallacy*, or the idea that moral motivation, moral responsibility, and moral judgement form an inseparable conceptual whole. I argued that the best way of dealing with the current theoretical impasse was by incorporating the notion of *moral subjecthood* (Rowlands 2011; 2012a; 2012b) into our discourse. The question we should be asking is not whether animals can be moral agents—that is, whether they can be held morally responsible for their actions—, but whether they are moral subjects; whether they ever behave on the basis of moral motivations. This chapter will argue that being a moral *subject* does not require mindreading capacities. In doing so, we shall return to Rowlands' characterisation of moral subjecthood and moral emotions, and show how his framework allows moral subjects to let go, not only of metacognition, but also of mindreading capacities.

Chapter 3 introduced the animal mindreading debate. I first went over the different reasons that philosophers have proposed to establish that only mindreaders can be moral. I then gave an overview of the experimental evidence that has been gathered throughout the last couple of decades and suggests the presence of some mindreading capacities in great apes and corvids. We saw that the results of these experiments could all be reinterpreted to fit the *behaviour reading* hypothesis (Heyes 1998; Povinelli and Eddy 1996; Povinelli and Vonk 2003; 2004; Penn and Povinelli 2007; 2013; Lurz 2009; 2011), where a behaviour reader is understood as an individual who can detect and read behavioural cues, as well as categorise them using abstractions, but who cannot postulate any underlying mental states with causal power over those behavioural cues. In this chapter I will show that, even if no animal were more than a behaviour reader, some of them could still behave on the basis of at least one type of moral emotion. In the process, I shall show that the arguments that philosophers have presented to account for a necessary connection between morality and mindreading are not successful.

Chapter 4 centred on the notion of empathy, and its connection to the concepts of morality and mindreading. While my survey of the literature was far from comprehensive, it sufficed to show that there is a pervasive ambiguity in the term 'empathy,' and that this ambiguity extends to its relation to the notion of mindreading. When empathy is understood as a mechanism or ability involved in *social cognition*, it is conceived either as a synonym for mindreading, as a mechanism that enables mindreading, or as an ability that presupposes mindreading. Within each of these different categorisations, empathy can be understood as related to perception-, theory-, or simulation-based accounts of mindreading. When empathy is thought of as a form of *moral motivation*, it can receive many different characterisations too, but scholars tend to agree that it will *require mindreading capacities.* To accommodate this theoretical commitment, the empirical evidence that points to empathy-based moral behaviour in animals is often reinterpreted using the aversive arousal hypothesis. In this chapter, I will argue that the dichotomy aversive-arousal/moral-motivation that lies at the heart of this reinterpretation is a false one.

I have been referring to moral behaviour, altruistic behaviour, and prosocial behaviour as though they were equivalent, given that this is a common trend in the literature on animal morality. However, it is now time to adopt a more precise vocabulary, since not all behaviour that is altruistic or prosocial is necessarily moral. Recall the ants who freed their entrapped conspecifics in the experiment by Nowbahari *et al.* (2009). Their behaviour was surely prosocial, since it benefitted others in the colony, but it would be hasty to label it moral solely on the basis of the external description. After all, a seatbelt may also save one's life in a car crash, but it would surely be extravagant to say that it is behaving morally. The same goes for altruistic behaviour. A bee may give her life to protect the hive by stinging an intruder. In doing so, she is behaving altruistically, for her behaviour benefits the members of her hive at a great cost to herself. To be able to describe the bee's altruism and the ant's prosociality as *moral* behaviour, however, something more is required. This something more can be found in the underlying motivations.

I'm going to understand that a certain behaviour is moral when it has been triggered by a motivation that is itself moral. This, of course, puts the *explanandum* in the *explanans*, but it is not an entirely vacuous definition. It means that looking at the external description of an individual's behaviour is not enough to be able to conclude that it is moral. For a behaviour to be moral, a sufficient, if not necessary, condition is that it has to have been triggered by a motivation of the appropriate sort. To understand this, imagine that we have two individuals, Tom and Jerry, who both decide to help an old lady carry her shopping bags. Imagine that Tom does it because he sees the lady struggling and feels it is his duty to help her, while Jerry does it solely because he sees it as an occasion to test his arm strength after a session at the gym. In this case, while the external description of the behaviour of both individuals would be equal, it seems safe to say that only Tom's behaviour qualifies as moral. Tom helps the old lady *for morally right reasons*, while Jerry is moved solely by selfish considerations.

I'm going to argue that one can behave in a way that is morally appropriate, and do so *for morally right reasons*, without possessing mindreading capacities. My strategy will be to offer a characterisation of a minimal moral motivation that I shall term *minimal moral empathy*. To develop my argument, I will divide this chapter into three parts. In the first part, I will go back to the arguments in favour of a necessary connection between morality and mindreading, which we saw in chapter 3, and using the example of *minimal moral empathy*, show how these scholars were mistaken. In the second part, I will offer a detailed definition of *minimal moral empathy*, go over its cognitive requirements, and attempt to account for its empirical plausibility. In the third and final part, I will give the reasons why *minimal moral empathy* should indeed be considered a moral emotion.

5.1. AN EMPATHIC BEHAVIOUR READER

As we saw in chapter 3, several important philosophical reasons point to a necessary connection between morality and mindreading. In particular, we looked at arguments presented by Cheryl Abbate (2014), Florian Cova (2013), Shaun Nichols (2001), and Jelle de Boer (2011), all of which seemed to suggest that moral behaviour necessarily requires the presence of mindreading capacities. In this section, I shall show that a closer look at these arguments reveals that they are not as convincing as they may initially seem. Along the way, I will present a characterisation of an empathic behaviour reader; an individual who lacks mindreading capacities, and yet behaves on the basis of a form of empathy that qualifies as a moral motivation. Instead of using merely the term 'empathy,' I will label the motivation behind this individual's behaviour '*minimal moral empathy*,' to illustrate the idea that my account distances itself from the common understanding of empathy as a motivation. As we saw in chapter 4, scholars usually understand that empathy is either not a moral motivation, or it is not minimal. In the latter case, empathy is thought to incorporate several intellectually demanding requirements, including mindreading. It has been traditionally established that empathy cannot be *both* moral *and* minimal. My aim throughout this chapter will be to show that Rowlands' framework allows us to combine the properties of morality and minimalism in a novel notion of empathy, which will be preliminarily introduced in this section, and then systematically defined in section 5.2.

5.1.1. ABBATE (2014): MINDREADING IS FOR RESPONSIBILITY

Let us begin, then, by looking at the arguments presented by Cheryl Abbate (2014). In this paper, her aim is to protect the moral patienthood of animals, that is, their status as legitimate objects of moral concern, from a potential threat that comes with higher-order thought (HOT) theories of consciousness. According to these theories, consciousness is gained through higher-order thought. Thus, one cannot feel pain unless one is capable of entertaining a thought to the effect that one is in pain. If HOT theories of consciousness were true, then animals would not be moral patients unless they were to be capable of entertaining higher-order thoughts, for their capacity to experience conscious pain and suffering (that is, their sentience) would be dependent on the former. Therefore, if HOT theories of consciousness were indeed true, those interested in safeguarding the moral patienthood of animals would have to prove that animals can indeed entertain higher-order thoughts. Abbate's worry is that succeeding in doing so might entail having to conclude that animals are not only moral patients, but moral agents too, for there seems to be a conceptual link between moral agency and higher-order thought. To sidestep this worry, she argues that moral agency requires a specific sort of higher-order thought—what she calls "moral thought third-order intentionality" (Abbate, 2014, p. 3)—and that possessing the sort of higher-order thought that confers sentience (and thus, moral patienthood) does not necessarily entail that one is a moral agent.

The reason why Abbate wants to argue against the idea that animals may be moral agents is because she considers that concluding otherwise may be detrimental to the case for animal rights/liberation. When offering the reasons why animals should be left out of the realm of moral agents, she states that the individuals that belong in this category must possess, on the one hand, "the capacity to form beliefs or thoughts about the mental states of others" (i.e. mindreading capacities) and, on the other hand, "the ability to assess one's own beliefs, desires, or thoughts about the mental states of others as right or wrong" (i.e. moral judgement capacities—what she calls 'moral thought third-order intentionality') (Abbate 2014, 9-10). To justify her claim that moral agency requires mindreading capacities, she argues the following:

> If one cannot understand that others suffer, feel pain, experience happiness, and so forth, then one cannot be held responsible for affecting another's mental states since one is unaware *that* one's actions affect others. (Abbate 2014, 10, her emphasis)

Her claim is, therefore, that mindreading capacities are necessary for moral responsibility. This is, of course, a perfectly coherent claim, stemming from the common link between understanding and moral responsibility. If Linda does not *know* that she is affecting others' mental states with her action, then it does not make much sense to hold her morally responsible for said action. And while there may be some room for doubt here (if, for instance, her lack of knowledge were due to a negligent ignorance of the relevant facts), it seems fairly uncontroversial that if Linda were to be entirely *incapable* of understanding that her actions affect others' mental states (due to a cognitive deficit of some form or another) then moral responsibility would almost certainly be out of the question.

While this idea is plausible, it does not affect the case I want to make. I will base my arguments on Rowlands' (2012b) distinction between moral subjecthood and moral agency—a distinction that Abbate (2014), perhaps due to the compelling nature of the *amalgamation fallacy* (see chapter 2), does not take into account, despite repeatedly quoting Rowlands (2012b). Instead, she misrepresents the concept of moral subjecthood as a way of referring to "a primitive form of morality" that might be found in animals (Abbate 2014, 21), akin to that hinted at by de Waal (see, for instance: de Waal 2006).[1] Rowlands explicitly denies such an interpretation of his ideas.[2] As the reader may recall, the category of moral subjecthood is not intended as a label for proto-moral creatures, but rather, as a label for those creatures (human or otherwise) who can behave on the basis of moral motivations (see chapter 2; Rowlands 2012b, chap. 3). Rowlands takes such motivations to include moral emotions, which encompass "sympathy, compassion, tolerance, and patience, and also negative or otherwise deficient counterparts like indifference, anger, malice, and spite" (Rowlands 2012b, 34). According to this author, if an animal and a human both behaved on the basis of, say, patience, they would both be moral subjects *in exactly the same sense* because the emotion that motivates their behaviour is moral *in exactly the same sense*.

Of course, there are many important differences between a (healthy adult) human and any animal. The human may reflect on the goodness or badness of the emotion she is experiencing. She may engage in ethical discussions with other humans. She may read

[1] Abbate (2014, 21) also misrepresents Bekoff and Pierce's work as defending this idea (see Bekoff and Pierce 2009, 10 for an explicit denial of this claim).

[2] "[A]nimals can act morally. There is no reason to think of their behavior as proto-moral—that is, quasi-moral in some rudimentary sense but falling short of the genuine article." (Rowlands 2012b, 33)

Spinoza's *Ethics*. But this will not make her more of a moral subject. These rational and reflexive capacities are relevant when it comes to determining the degree of understanding that a human can have when deciding to behave in this way or another. The degree of understanding an individual is capable of will be relevant in order to decide how responsible she is for her actions, that is, whether and to what extent she is a moral agent. Thus, the animal and the human may both be motivated to behave by patience, but (presumably) only the human possesses a sufficient understanding of her behaviour and the surrounding circumstances to merit being held responsible, and ultimately, praised for what she does.

With these considerations in mind, we can go back to Abbate's argument and see that, while it is probable that mindreading capacities are necessary when it comes to being held responsible for one's behaviour, this would only mean that moral *agency* requires mindreading capacities. The claim that I'm interested in defending is that being a moral *subject* does not require mindreading capacities. Thus, Abbate's argument does not affect my case. Furthermore, considering the question of whether animals can be moral, not as the question of whether they can be moral agents, but as the question of whether they can be moral subjects, allows us to see that Abbate's fear towards the possibility that animals may be moral is unfounded. She is worried that concluding, on the basis of the empirical evidence, that animals are moral creatures would be "contrary and detrimental to the animal ethics vision" because it may:

> open the door to clichés such as the claim that human beings are justified in eating meat or exploiting nonhuman animals because they "deserve" such treatment since they themselves viciously and cruelly kill, hurt, and injure other animals (Abbate 2014, 22-3)

Abbate is right in pointing out that defenders of animal morality often offer a sugar-coated vision of animal behaviour, and that opening the door to the possibility that animals may be motivated to act by moral emotions means, not only considering that they may be motivated by empathy or compassion, but also by their negative counterparts (Abbate 2014, 22). This would, however, only justify the use of an argument from desert to validate the exploitation or killing of animals if we were to be talking about animals as moral agents—and even then it would amount to a poor argument. Mere moral subjects (presumably the most that an animal can aspire to) cannot be held responsible for their behaviour, which means they cannot be legitimately punished for it. There is, therefore, no reason to worry in this respect.

5.1.2. Cova (2013): Mindreading is for caring

Cova (2013) distinguishes two positions in the animal morality debate: (1) *continuism*, or the idea that some of the capacities that make humans moral are present, if somewhat rudimentarily, in some nonhuman species, and (2) *discontinuism*, or the thesis that only humans posses these capacities. In a similar vein as Rowlands, Cova argues that in order to make progress in debates on animal morality, we need to introduce modifications into the way we use our moral terms. However, in contrast to Rowlands' proposal of separating moral motivation from moral responsibility, Cova (2013) argues that the disagreement amongst continuists and discontinuists can be at least partially solved by separating moral judgement from moral agency. He understands moral agency as the quality of being "morally responsible of (some of) [one's] action" and moral judgement as the ability to "judge whether something is right or wrong" (Cova 2013, 118). The key to solving the debate, according to Cova, is realising that one can be a continuist about moral agency and a discontinuist about moral judgement, and thus, that we can establish that "we share the psychological bases for moral agency with other animals" while acknowledging that "we are the only known species able to form moral judgments" (Cova 2013, 118).

In order to defend his position, Cova attempts to construct an account of moral agency without moral judgement. To this end, he gives a series of examples of people performing good or bad deeds without stopping to reflect on their goodness or badness, and argues that it is counterintuitive to consider that the agents were not morally responsible for these actions just because they didn't execute their moral judgement capacities while performing them. It seems, however, that he is not properly addressing the issue at hand, because in all the cases he considers the agents are healthy adult humans who *do* have moral judgement capacities. As a result, they can understand (even if they haven't stopped to reflect upon it in these particular instances) that what they're doing is right or wrong, and that they should be praised or blamed for it. Cova does not succeed in showing that this isn't part of the reason why their actions are worthy of praise or blame. For example, Cova considers the following case:

> Let's say that Jack had a bad day, was irritated, and smashed the window [*sc*. of the car parked in front of him] without taking the time to assess whether it was right or wrong. Let's also say that, though he realized afterwards that it was the wrong thing to do, he did not regret this action at great length. Should we say that Jack is not responsible and does not deserve blame for what he did? That he hasn't the duty to pay for repairs? That seems very counter-intuitive. (Cova 2013, 123)

An obvious response would be to say that the fact that Jack has moral judgement capacities, but has failed to make proper use of them, is an important part of what warrants our holding him morally responsible for his action. The fact that he has these capacities does, indeed, make it counterintuitive to say that he is not responsible. On the other hand, if Jack were a severely disabled individual who lacked all capacity to understand that what he did was wrong, then holding him responsible or making him pay for repairs would also be very counterintuitive. Cova does not put forward any argument to prevent us from interpreting his examples this way.

Cova's characterisation of a 'mere' moral agent is also a rather awkward one. He states that a moral agent that were incapable of engaging in moral judgements would be "responsible for his action" but that this would not "necessarily entail that he can be punished," because "one has to be both a moral agent and a moral judge to be an appropriate target of punishment" (Cova, 2013, p. 129). Cova defends this last claim by appealing to his own intuition ("it seems to me that we want the people we punish to understand why they are punished"), and also to an experimental study by Gollwitzler and Denzler (2009) that suggests that people "consider revenge satisfactory only if the offender understands (and acknowledges) that revenge was taken against him because and in virtue of a prior unfair behaviour" (Cova 2013, 129). However, unless we speak of degrees of moral responsibility (which Cova does not), it is very difficult to see how someone can fulfil the cognitive requirements for being held morally responsible for an offense and for this not to entail that she can be punished. It seems that any consideration that made reference to the individual's personal characteristics (and not to circumstantial facts) and served to undermine the legitimacy of punishment would also entail that the person in question is not a full-blown moral agent. The whole point of holding someone *morally* responsible is to open the possibility of praising or blaming her, and rewarding or punishing her if the circumstances warrant it. Cova's appeal to the intuition that those we punish need to understand why they are punished only works against his own interest, because it points, precisely, to the fact that moral responsibility and moral judgement cannot be separated.

We have seen that Cova's argumentative strategy to defend the logical independence of moral agency and moral judgement is unsatisfactory, and also, that his characterisation of a 'mere' moral agent is rather counterintuitive. However, the most important point that needs to be made here is a more fundamental one, namely, that Cova *does not need* to separate moral agency from moral judgement. He does not need to do so, because he is merely trying to make sense of behaviours that can be *morally motivated*, that occur "for good reasons" (Cova 2013, 127), in the absence of moral judgement capabilities. Appealing to the category of moral subjecthood is a much easier way of achieving this. An individual who can behave on the basis of morally good reasons (or morally bad ones) is a moral subject. If that subject *also* possesses the ability to engage in moral judgements, then she is *also* a moral agent, and can be held responsible (and, thus, praised or blamed) for her behaviour. This allows us to make sense of the morality of the examples Cova uses even if we suppose that the subjects involved lack all capacity to engage in moral judgements. At the same time, we can also avoid the awkward consequences that come from separating moral agency and moral judgement.

I have a further concern with Cova's paper, which will bring us to the issue of mindreading. It seems to me that he doesn't fully succeed in separating moral behaviour from moral judgement. To see why this is so, let us consider one of his main aims in this paper, which is to give a minimal account of what it means to act for morally good reasons. Cova reaches the conclusion that "one acts for the good reasons when one actually cares about the person one is trying to help" (Cova 2013, 127). In order to be able to say that an individual *cares* for another, two conditions must be met, according to Cova: (1) she must be "able to understand that this person has *interests*," and (2) she must "give an intrinsic (i.e. non-instrumental) positive value to the fact that this person's interests are preserved and augmented" (Cova 2013, 127, his emphasis). These two conditions, in his view, mean that being a moral creature only requires "[acting] according to what we attach importance to and a bit of theory of mind [i.e. mindreading capacities]" (Cova 2013, 128).

Cova, therefore, considers that mindreading capacities must be present in order to be able to say that an individual *cares* for others, and thus that she helps others for morally good reasons. My reason for thinking that Cova does not succeed in letting go of moral judgement capacities resides in the other condition he puts forward, namely, the idea that the moral individual must also be capable of *attaching importance* or *giving an intrinsic positive value* to the preservation of others' interests. Attaching importance or giving a positive value to X seems to mean, precisely, (consciously or unconsciously) judging that X is a good thing. How is this different from moral judgement? How can a person attach importance to something without the ability to engage in moral judgements? Cova does not offer a solution to this question, which means it poses a potential threat to his whole argument. I believe, however, that there is a way around this problem; a way that will allow us to effectively separate moral behaviour from moral judgement, and additionally, from mindreading capacities.

My proposal is the following. We can construct a minimal account of care that will not require the ability to (consciously or unconsciously) judge that preserving someone's interests is a good thing, by focusing on the emotions that an individual might undergo when deciding to help someone. So, let's imagine an individual called Higgins who feels sad whenever he sees others in distress. Let's imagine that this sadness is what triggers Higgins' urge to help or comfort them, and that he feels happy once they're no longer in distress. Even if Higgins were incapable of engaging in moral judgements, Rowlands' (2012b) framework allows us to categorise Higgins' behaviour as moral because it is motivated by an emotion that *tracks* the moral proposition "This creature's distress is bad." An emotion can be said to track a moral proposition if there is a truth-preserving relation between the emotion and the proposition in question, so that the truth of the proposition is guaranteed whenever the emotion is not misguided (see sections 2.3.2 and 5.3.2; also Rowlands 2012b, chaps. 2, 9).

To see what this means, imagine that Higgins is sad because he sees that Jane is crying, and that his sadness is *intentionally* directed at Jane's crying, that is, Higgins is sad *that* Jane is crying. His being sad that Jane is crying means he is *experiencing* Jane's crying as something unpleasant; as something bad. Built into his sadness at Jane's crying is an urge to comfort her. This experiential form that Higgins' emotion takes is what allows us to speak of it tracking the proposition "This creature's distress is bad" (see also section 5.3.2). Now, suppose that Jane's crying were due to an entirely mundane reason, such as the fact that she's watching a sad film. Or suppose that Jane had committed a serious offense and her crying was the result of her well-deserved punishment. In these cases, Jane's crying would (at least arguably) *not* be a bad thing. Higgins' sadness would then be *misguided*, because he would be experiencing as bad something that is not, in fact, a bad thing. This is what it means to say that the truth of the proposition "This creature's distress is bad" is guaranteed by the non-misguided status of Higgins' emotion.

The idea of emotions that track moral propositions allows us to separate moral behaviour from moral judgement entirely. We do not need Higgins to be capable of entertaining a proposition such as "This creature's distress is bad." Higgins can lack all capacity to engage in moral judgements *and still be a moral subject*, because the morality of his behaviour comes by virtue of the fact that it is triggered by an emotion that tracks a moral proposition. And here comes the crucial bit—since we can let go of all capacity to entertain the proposition "This creature's distress is bad," we can also let go of all capacity to entertain the proposition "This creature is distressed." That is, we do not need Higgins to be capable of attributing any mental states to others, because what is important is that he reliably undergoes sadness when witnessing someone's distress. It is enough for Higgins to be a *behaviour reader*, and for his sadness to be triggered by, and intentionally directed at, the *superficial behavioural cues* that accompany Jane's distress—we do not need him to understand anything about the underlying mental states.

For Higgins to be a minimal moral subject, we need his behaviour to be motivated, at times, by a (minimal) moral emotion. This (minimal) moral emotion can be of different sorts. In the example I have been using, Higgins is motivated by the emotion I shall call *minimal moral empathy* (see section 5.2 for a definition of this notion). For this specific kind of moral emotion to obtain, Higgins must possess a reliable mechanism that ensures that, upon detection of the superficial markers of distress in others, he experiences an emotional contagion that results in a form of distress intentionally directed at the state of affairs that is the other individual in distress. We need the resulting emotion to be what moves Higgins to help or comfort the other. As we shall see, fulfilling these conditions is enough for Higgins to count as a moral subject, regardless of whether or not he is a moral judge, a mindreader, or indeed, a human being.

5.1.3. NICHOLS (2001): MINDREADING IS FOR MOTIVATION

In this paper, Nichols gives a minimal account of the cognitive mechanisms that underlie what he considers one of the "basic moral capacities: the capacity for altruistic motivation" (Nichols 2001, 425). He focuses on the core cases of human altruistic motivation[3]—namely, "cases of helping or comforting others in distress" that emerge in early childhood and are pervasive among adults (Nichols 2001, 428)—and argues that they can be best accounted for by postulating the existence of what he calls a 'Concern Mechanism' (Nichols 2001, 426). This is an affective system that produces the motivation to engage in helping or comforting behaviour when we encounter someone in distress. One of Nichols' aims in this paper is to determine what kind of mindreading mechanisms must be in place for the 'Concern Mechanism' to be activated. He argues against two different options: (1) the possibility that all altruistic motivation may be triggered by perspective-taking processes, which he finds too intellectually demanding and empirically implausible (Nichols 2001, 440-3), and (2) the possibility that no mindreading capacities at all are required for altruistic motivation to occur. It is in his arguments against option (2) where we find a defence of the idea that mindreading capacities are necessary for moral behaviour.

Minimal moral empathy (MME) would be an example of what Nichols considers the most "radical" view of the relationship between mindreading and altruistic motivation, namely, the view that no mindreading at all is required for altruistically motivated behaviour to occur (Nichols 2001, 426). The core cases of altruistic motivation, according to Nichols, cannot be adequately accounted for from this 'radical' view because without mindreading one will not be motivated to help another when escaping is easier. In particular, Nichols considers that we need at least a "minimal mindreading capacity to attribute negative affective or hedonic states to

[3] Nichols claims to be giving a minimal account of *altruistic* motivation, but I will understand his paper as giving a minimal account of *moral* motivation. Even though I said at the beginning of this chapter that morality and altruism are not the same, this treatment of Nichols' paper is warranted by his claim that altruistic motivation is one of the "basic moral capacities" (Nichols 2001, 425), as well as by the fact that he speaks of altruistic *motivation* and not merely altruistic *behaviour*. While altruistic behaviour in and of itself isn't necessarily moral, it becomes so when it has been triggered by an altruistic motivation.

others" in order for altruistic motivation to take place (Nichols 2001, 346).

Continuing with our example, Nichols would say that the lack of mindreading capacities will mean that Higgins, due to his emotional contagion, experiences Jane's distress as "bad music" that he'd like to turn off, and escaping will be just as good a solution as comforting her (Nichols 2001, 429). Possessing a capacity to represent negative affective states, on the other hand, would mean that "escape is not an adequate alternative" because "the motivation comes from an enduring *internal* cause" (Nichols 2001, 435, his emphasis). If Higgins possessed mindreading capacities, his motivation to help Jane would not have been triggered by mere superficial cues, but by a representation of her distress, which would mean that "merely escaping the perceptual cues of pain won't eliminate the consequences of the enduring representation that another is in pain" (Nichols 2001, 436).

The most obvious response here would be to argue that MME can also take the form of an "enduring internal cause." We are not supposing that Higgins' emotional contagion results in self-directed personal distress, or in a form of distress with no intentional object, but rather, that as a result of his emotional contagion, Higgins is distressed *that* Jane is displaying distress behaviour. Jane's distress behaviour is the intentional object of Higgins' distress—he experiences it *emotionally* as distressing. If the emotional contagion gives rise to an emotion that is *intentionally directed* at Jane's distress behaviour, then the cause of Higgins' distress can persist in his mind even after he escapes the situation. Nichols considers this possibility but dismisses it:

> An emotional contagion theorist might continue to deny any role for mindreading and maintain that altruistic motivation comes from an enduring representation of the behavioral, acoustic, or physiognomic cues that cause emotional contagion ... The problem is that ... if one knows that the cues leading to emotional contagion are merely superficial, this typically does not prevent one from experiencing emotional contagion, but it does undermine altruistic motivation. (Nichols 2001, 435)

This response is unsatisfactory because it presupposes that the being in question possesses mindreading capacities and can "[know] that the cues leading to emotional contagion are merely superficial." If an individual were to lack mindreading capacities altogether, then she wouldn't be capable of distinguishing "merely superficial" cues from those that are markers of underlying mental states, and so there is no reason to suppose that the superficial cues couldn't provide the adequate motivation. All that is required is for the individual's past learning experiences or hardwired behavioural tendencies to move him to engage in affiliative behaviour as a result of his (intentional) emotional contagion (see section 5.2.4).

In contrast to an account of altruistic motivation based on emotional contagion, Nichols proposes that the attribution of distress to others may be precisely what triggers the affective process that results in a motivation to behave altruistically. This is what Nichols terms the 'Concern Mechanism,' which he proposes as underlying much of our altruistic behaviour, and would work as follows:

> The distress attribution might produce a kind of second order empathic distress in the subject. For example, representing the sorrow of the target might lead one to feel sorrow. This would provide a kind of empathic motivation for helping. And the motivation would be effective even when escape is easy. For the cause of the emotion is still the representation of the other's mental state and as a result, one is motivated not simply to escape the situation since that would not rid one of the representation. (Nichols 2001, 445-6)

We have seen that there is no reason to suppose that emotional contagion cannot trigger an enduring representation. At the same time, there is no reason to suppose that "escaping the situation" cannot rid one of the representation of another's distress. We have all experienced being overwhelmed by a particularly tragic story in the news and changing channel to avoid being put off our dinner. There need not be any images of people in distress that could trigger an emotional contagion for this to occur. It could simply be the case that a journalist is reporting a story that makes us very aware of the suffering of the people involved in it. By changing channel, we become distracted and thus stop thinking about their distress—that is to say, our internal representation of it is 'turned off.' Of course, the cases that Nichols is thinking about are those in which we are faced with the dilemma of helping the person in distress or leaving the scenario—in this respect my example is not a perfect analogy. Nevertheless, it serves to illustrate the fact that having an internal representation of someone's distress does not mean that it must endure once we escape the situation—we can, after all, make the effort to become distracted and think about something else.

Having mindreading capacities does not ensure that one will help when escaping is easier. The 'Concern Mechanism' that Nichols proposes would therefore not necessarily be infallible. The same applies to MME, about which we can never say that it will *necessarily* result in comforting or helping behaviour—there could always be other intervening factors that motivate the individual towards a different behavioural outcome. For both Nichols' 'Concern Mechanism' and MME to result in helping or comforting behaviour, a number of further conditions must probably obtain: the individual must have the right beliefs, her relation to the distressed person must be of a certain sort, there must not be any other pressing issue to attend to, and so on. This is not a problem for my proposal, since I am not attempting a description of an infallible moral motivation. In fact, such an enterprise would probably be doomed to failure, since motivations never occur in isolation, but in individuals who have many other mental states and behavioural dispositions, so it is unlikely that a motivation can ever guarantee a certain behavioural outcome.

It is important to emphasise that Nichols is directing his arguments at the thesis that *all* altruistic motivation is caused by emotional contagion, and this is not what I'm interested in defending. It certainly seems plausible to me that something akin to Nichols' 'Concern Mechanism' may be at the basis of much human (perhaps even nonhuman) altruistic behaviour.

My point is that MME may *also* be a kind of altruistic motivation and that the arguments put forward by Nichols don't succeed in proving otherwise. Of course, whether MME ever does take place in nature is an empirical matter on which I take no stand (although I do believe that it is plausibly the case—see section 5.2). My interest is in arguing that, if an individual (at times) behaved on the basis of moral emotions such as MME, then she would be a moral subject, regardless of whether or not she possessed mindreading capacities.

5.1.4. DE BOER (2011): MINDREADING IS FOR RECIPROCITY

Jelle de Boer has also attempted to give an account of the "basics of morality" (de Boer 2011, 893). He claims to be unconvinced by the anecdotes that primatologists, such as Frans de Waal, have put forward as evidence of moral behaviour in great apes. In his view, observations of apes being helpful towards conspecifics are not enough to conclude that they are moral creatures. Their underlying motives and mechanisms must be of an appropriate sort for them to warrant this label. De Boer thus subscribes the idea that the merely external description of an individual's behaviour cannot conclusively prove its moral character. Accordingly, de Boer proposes a series of cognitive requirements that a moral individual must meet. These conditions function as steps that will take us from the "sometimes-helpful chimpanzee" to the "moral man" (de Boer 2011, 892). Upon witnessing chimpanzee Jakie helping chimpanzee Krom, we could conclude that Jakie is a moral creature if we knew that the following three conditions obtained:

> Jakie believes that he should help Krom.
> Jakie believes that he should help Krom because Krom needs help.
> Jakie believes that everybody believes that he should help Krom because Krom needs help. (de Boer 2011, 903)

For ease of exposition, let us now adapt these three conditions to the example we have been using all along. For Higgins' comforting behaviour towards Jane to be moral, de Boer would require that Higgins fulfil the following conditions:

1. Higgins believes that he should comfort Jane.

2. Higgins believes that he should comfort Jane because Jane needs comfort.

3. Higgins believes that everybody believes that he should comfort Jane because Jane needs comfort.

Condition (1) is what de Boer calls the "Belief" condition. It is not enough for Higgins to comfort Jane on the basis of "biochemistry" or of "stimulus response," in the way that ants automatically respond to the chemical signals from their conspecifics. The basis for Higgins' affiliative behaviour towards Jane must be the belief that he *should* comfort Jane (de Boer 2011, 895). While the belief that one should help or comfort another is a perfectly valid moral motivation, it seems far-fetched to hold that it is *the only form* that a valid moral motivation can take, as de Boer seems to be implying. Surely, there must be some middle ground between behaving on the basis of a biochemical stimulus, and entertaining such an intellectually demanding belief as "I should help X." If we are trying to determine the "basics" of morality, we should aim to keep things as minimal as possible. MME as I have described it would be a different kind of moral motivation, and a less intellectually demanding one at that. If Higgins helps Jane on the basis of MME, his behaviour would be moral even if he cannot entertain a proposition such as "I should comfort Jane," because it would have been triggered by an emotion that *tracks* this proposition.[4] We have already seen that this tracking relation allows us

[4] I use this proposition, instead of "This creature's distress is bad," as I have been doing until now, to keep in

to make sense of the morality of a motivation in the absence of moral judgement capacities (which would be necessary to entertain the proposition "I should comfort Jane"), so let us move on to the other two conditions put forward by de Boer.

Condition (2) is what this author calls the "Reasons" condition. According to it, Higgins has to be motivated to comfort Jane by the belief that "Jane should be comforted," and the *reason* for believing this must be the fact that Jane needs comfort. If Higgins were to believe that Jane should be comforted because that would give Higgins access to "social approval and material benefits," then Higgins would be behaving on the basis of the "wrong kind of reasons" (de Boer 2011, 898), because his ulterior motive would be a selfish one. The right kind of reason for thinking that Jane should be comforted is, according to de Boer, precisely the fact that Jane *needs* comfort. Now, of course, Higgins, upon witnessing Jane's distress behaviour, cannot *understand* that Jane needs comfort unless he possesses mindreading capacities. If Higgins were a mere behaviour reader, he could not fulfil condition (2).

However, the fact that Higgins is behaving on the basis of an emotion with content—he is motivated to comfort Jane because he experiences her distress behaviour emotionally as distressing—makes the motivation behind his comforting behaviour more than a mere cause; something close to a reason (see section 5.3.4). Higgins' comforting behaviour is not simply triggered automatically by his MME. Rather, the content of this emotion and the experiential form it takes is what motivates him to comfort Jane. Even though Higgins cannot reflect upon it, his finding Jane's distress behaviour distressing is the reason behind his comforting behaviour, and this justifies his action from the point of view of an external observer. We, as moral agents with moral judgement, can determine that Higgins comforts Jane *for the right reasons*. He is

line with de Boer's arguments. The specific proposition that an emotion tracks will be determined by the latter's phenomenal character (see section 5.3.2.). In the case of MME, since there is an urge to comfort Jane that is built into Higgins' emotional contagion, we could also describe it as tracking the proposition "I should comfort Jane." Both propositions would be false if Higgins' behaviour were to be misguided.

feeling what he *should* feel, given the circumstances. Since we are not discussing whether Higgins should be praised or blamed for his behaviour, but merely whether the motivation behind it is moral, it sufficient to acknowledge that for us, as external observers with moral judgement capacities, experiencing Jane's distress behaviour as distressing is a morally appropriate reason for comforting her. The fact that the evaluation is external, rather than internal, does not make Higgins' motivation any less moral—although it would certainly make him less worthy of praise.

Those fond of the aversive arousal hypothesis (see section 4.3) might not be satisfied with this line of reasoning. Higgins, they might argue, is acting out of a purely selfish desire to end his own distress at Jane's distress. His concern is solely for himself, and thus there is no reason to label it moral. I do not wish to deny that MME bears a strong resemblance to aversive arousal, insofar as the key idea here is the unpleasantness that Higgins experiences upon witnessing Jane's distress behaviour, and which he does indeed aim to eliminate via the extinction of its cause. But it would be a *non sequitur* to conclude that this makes Higgins' reaction selfish. In the following extract from Rowlands (2012b), we can see why this is so:

> [C]onsider the probably apocryphal story of Abraham Lincoln, who ordered his carriage driver to stop and rescue two young birds that he spied were in distress. On being complimented for his kindness, he responded that it was, in fact, purely a matter of self-interest: He would not have been able to sleep at night thinking about those birds. There is a well-known error embodied in Lincoln's response. Why would the plight of the birds ruin Lincoln's sleep? It would do so if Lincoln were, in some respects at least, a compassionate person. If he were callous or indifferent, then the birds' suffering would have no impact on his slumbers. (Rowlands 2012b, 10)

Similarly, we could say, if Higgins were a callous or indifferent individual, he would not find Jane's distress behaviour unpleasant. We know that Higgins is an *empathic* individual because he has a reliable disposition to experience MME upon seeing someone in distress. If Higgins, upon seeing others in distress, experienced an emotional contagion that reliably resulted in an urge to escape the situation, we would say that his emotion tracks the proposition "*My* distress is bad," rather than "This creature's distress is bad." In a case like this, I think it would be fair to label it a selfish reaction. MME, we could say, is a special form of aversive arousal, insofar as (1) the intentional object of Higgins' distress is not his own distress, but Jane's distress behaviour—this is what Higgins is distressed *about*—, and (2) built into it is an urge, not to escape the situation, but to comfort Jane. This, as I will argue more thoroughly in upcoming sections, is enough to conclude that Higgins' motivation is a *moral* one.

De Boer's condition (3) is the "Mutual beliefs" condition. Higgins is required to believe that everybody believes that he should comfort Jane. The rationale behind this condition is twofold. Firstly, it is meant to account for the fact that Higgins should be "sensitive to the relevant beliefs" of Jane; that her attitude and circumstances should count. De Boer's view is that helping another "without a concern for what the other party thinks about the issue, seems dogmatic, disconnected" (de Boer 2011, 900-1). Perhaps Jane, in these particular circumstances, does not want to be comforted. Perhaps she just wants to be left alone. Requiring Higgins to believe that everyone believes that she should be comforted is meant to secure the objective character of the appropriateness of Higgins' action. And of course, requiring Higgins to be capable of believing that everyone believes... means requiring that he possesses mindreading capacities. Additionally, without these capacities, Higgins cannot be sensitive to the particular nuances of Jane's situation.

Once again, de Boer seems to be asking too much of the "basic" moral individual. Would we not want to say that Higgins' motivation to comfort Jane would be moral, even if Jane in those circumstances did not want or did not need to be comforted? If Jane were displaying distress behaviour merely because she is rehearsing her part in a play, Higgins' attempt to comfort her would be misplaced, because in these circumstances, Jane is displaying feigned and not 'real' distress behaviour. But would Higgins' behaviour not be moral, nevertheless? To see this, consider an analogous case. Something gets in your eye and you start tearing up. Upon seeing this, your friend Sally comes up to you in a genuine attempt to comfort you. Would you not be inclined to think that Sally is a caring person, even though you don't need to be comforted? In this example, Sally's behaviour is also based on an emotion that is misplaced because it is grounded in a false factual assertion, i.e., she is interpreting your tears as a sign that you are distressed when, in fact, you're not. While Sally would, of course, deserve praise in a way that Higgins could not, the example suffices to show that, *pace* de Boer, behaviour based on emotions that have misfired (either because they are misplaced or because they are misguided) can still be moral. And the tracking account provided by Rowlands allows us to preserve this characteristic of moral behaviour (see also section 5.3.2).

The second motivation behind de Boer's condition (3) is that it allows the reason behind Higgins' comforting behaviour to acquire the status of a moral norm:

> [I]n a moral world it is not so that each individual simply believes what should be done, each on his own, with all merely converging on the moral truth. What should be done is a public fact in a moral society: everybody believes this about each other. This 'everyone believes that everyone, whatever the role he happens to occupy, believes what should be done' establishes a moral norm. (de Boer 2011, 902)

If everyone in a society believes that people in distress who need to be comforted *should* be comforted, then this becomes a moral norm of the society. If Higgins believes that everyone believes this (which requires mindreading), then he understands his reason to comfort Jane as a moral norm, and this, according to de Boer, would turn him into a full-blown moral creature. This reciprocity, this everyone believing this about everyone else, turns the proposition "This creature in distress should be comforted" from a personal reason into a moral norm of the society, which allows there to be a mutual accountability. Whoever does not behave in accordance with this norm can be held responsible, for it would now have the status of a moral fault.

This would be a reasonable requirement if we were attempting to give an account of norms, both moral and social. It makes *prima facie* sense to require everyone in a society to believe that everyone else believes that a certain norm obtains, in order for this norm to obtain (although, see Andrews 2009 for an account of norms without mindreading). If we're trying to give an account of moral motivations more generally, however, there is no reason why their existence can't often be determined at the personal level, especially if we adopt a pluralistic standpoint—as I've been doing—and consider that moral norms are not the only type of moral motivation that exists, but that one can also behave morally on the basis of moral emotions, character traits, or moral reflections that may go against what is generally held as the moral truth. In a very primitive human society where no moral norms existed, we can nevertheless imagine individuals being cruel or envious, kind-hearted or compassionate. By acknowledging the existence of different kinds of moral motivations, we open up the possibility that individuals in a society that hasn't developed moral norms may nevertheless behave morally. De Boer's framework is ultimately unconvincing because it excludes this possibility.

5.2. DEFINING *MINIMAL MORAL EMPATHY*

In the previous section, we have seen that the different reasons that philosophers have put forward to account for a necessary connection between morality and mindreading can all be questioned. My argumentative strategy in my attempt to refute them has been to introduce a hypothetical moral emotion—what I have termed *minimal moral empathy* or MME. I have tried to show that this notion is theoretically conceivable and that it enables us to make sense of moral behaviour in the absence of mindreading capacities. In this section, I intend to show that this hypothetical emotion is, in fact, empirically plausible. To that end, I shall analyse its cognitive requirements and present the evidence we have of their presence in nonhuman species.

My point of departure must first be to offer a more systematic characterisation of MME. We can define it as follows:

> *Creature C possesses* minimal moral empathy *(MME) if: (1) C has an ability to detect distress behaviour in others, and (2) due to the action of a reliable mechanism, the detection of distress behaviour in others results in a process of emotional contagion that (3) generates a form of distress with the other's distress behaviour as its intentional object, built into which is (4) an urge to engage in other-directed affiliative behaviour.*

Note that the four conditions offered are understood as a set of *sufficient* conditions for the possession of a minimal form of empathy as a moral emotion. That is, if C fulfils these four conditions, then that is enough to determine that she possesses MME, and thus, that she is a moral subject (see also section 5.3). However, these are not necessary conditions, so other minimal notions of empathy as a moral emotion are perhaps conceivable. It is not necessary for my purposes to take a stand on this issue. My interest in this section is to argue that this particular conception of empathy as a minimal moral emotion is empirically plausible. Let us now have a closer look at each of the four conditions that conform MME.

5.2.1. AN ABILITY TO DETECT DISTRESS BEHAVIOUR

As we saw in chapter 3, research on mindreading capacities in animals has focused on the ability of different species to attribute *cognitive* states to others, in particular, perceptions, intentions, beliefs, and states of knowledge/ignorance. Animals' capacity to attribute *affective* states, such as distress, remained unaddressed for the most part of the 20th century, despite the fact that affective states are also mental states. An exception to this are the studies conducted in the fifties and sixties by Robert Miller *et al.*, who found that rhesus monkeys could distinguish fearful from non-fearful expressions in the faces of conspecifics, and use the former as cues for pressing a lever that would terminate an imminent shock (Miller, Murphy, and Mirsky 1959; Miller, Banks, and Ogawa 1963; Miller, Caul, and Mirsky 1967). An isolated study by Gary Berntson *et al*. (1989) also determined that the heart rate of chimpanzees will decelerate in response to recordings of conspecifics screaming, and accelerate in response to recordings of chimpanzee laughter. These studies provided preliminary evidence of an ability to attribute affective states in these two species.

It wasn't until the turn of the century that scientists began to gather further evidence for animals' capacity to attribute affective states to others. Most of this evidence corresponds to primate species. For instance, thanks to the research performed by Lisa Parr and colleagues, we now know that facial expressions are a very important part of primate social cognition (Parr et al. 2000; Parr 2001; Parr 2003; Parr and Heintz 2009). Parr has shown that chimpanzees can discriminate between five different types of facial expressions in conspecifics, all of which correspond to different affective states, and that this discrimination can occur even when chimpanzees are unfamiliar with the individual (Parr, Hopkins, and de Waal 1998). This capacity appears to be present as well in crested macaques (Micheletta et al. 2015). Parr has also demonstrated that chimpanzees are capable of matching photographs of stimuli with positive or negative emotional valence (such as, correspondingly, toys and hypodermic needles) to photographs of conspecifics displaying facial expressions with similar emotional valence (Parr 2001).

More evidence of primates' capacity to attribute affective states to others comes from a field playback experiment performed by Katie Slocombe, Simon Townsend, and Klaus Zuberbühler (2009), which found that members of this species can discriminate between those screams from unrelated conspecifics that are vocalised in response to severe aggression, and those that are vocalised in response to mild aggression or tantrums. Yo Morimoto and Kazuo

Fujita (2012) further found that Capuchin monkeys are capable of extracting information from the emotional expressions of conspecifics. After witnessing a conspecific's emotional reaction to the contents of two different containers (either food or a stuffed toy), the experimental subjects were allowed to choose one of the two containers, and were found to reliably pick the one that their companion had had a positive emotional reaction to.

With respect to non-primate species, the attribution of affective states to conspecifics has only just begun to be studied, so there is still a very limited amount of evidence. An example of this is an observational study by Joshua Plotnik and Frans de Waal (2014), who found some preliminary evidence of a capacity to discriminate distress in others in Asian elephants. These were found to engage in affiliative behaviours and vocalisations towards conspecifics more often when the latter were displaying distress behaviour, than during control periods, and these affiliative interactions were usually unsolicited, which suggests that elephants can discriminate when conspecifics are distressed.

An experimental study by Satoshi Nakashima *et al.* (2015) found preliminary evidence of laboratory rats' capacity to discriminate when conspecifics are displaying distress behaviour. The experimental subjects were placed in an apparatus that consisted of three zones: a centre zone (where they were placed), a compartment to the right that had images of conspecifics displaying pain behaviour, and a compartment to the left that had images of conspecifics displaying neutral behaviour. The rats were left to explore the apparatus for several minutes, and the experimenters found that they preferred to spend time in the compartment with images of neutral behaviour, which suggests that they recognised the images of pain behaviour as such and were intentionally avoiding them.

There is also some evidence concerning inter-species communication of affects. David Buttelman, Josep Call, and Michael Tomasello (2009) investigated whether members of the four species of nonhuman great apes (chimpanzees, bonobos, orang-utans, and gorillas) would be able to use emotional expressions from humans as cues to find preferred food. The apes saw a human actor react either neutrally, happily, or with disgust to the food contents of two different boxes. The apes were then left to choose a box. In the Happy-Disgust condition, the apes reliably chose the box that the experimenter had reacted happily to. In the Happy-Neutral condition, no difference in performance was observed. The experimenters interpreted this as a possible sign that “apes distinguish between human emotional expressions when they are very distinct, but cannot distinguish expressions that are more similar” (Buttelmann, Call, and Tomasello 2009, 692). In a second experiment, the apes saw the human react with an expression of either happiness or disgust to the contents of two different boxes, and then eat something. The apes were then given a choice of containers, to see if they would be able to infer that the human had eaten the food towards which he had reacted positively, and, so, that they should choose the other box, which would presumably still have food in it. The apes were found to perform correctly in this test at above chance level, which means, according to the experimenters, that they were able to use the human’s “emotional reactions to things to predict his behaviour, even in the absence of any other behavioural cues” (Buttelmann, Call, and Tomasello 2009, 696).

Susan Lingle and Tobias Riede (2014) recently used a field playback experiment to test the response of mule deer to distress vocalisations of members of different species, including sea lions, seals, marmots, humans, bats, cats, and dogs. Female deer were found to approach the speaker whenever it was playing recordings of vocalisations from these different species “as though they were going to assist a fawn in distress,” which was interpreted as evidence that distress vocalisations of different mammals share acoustic traits, and that mule deer have a sensitivity to these traits (Lingle and Riede 2014, 517).

The remaining evidence of inter-species affective mindreading comes from studies on domestic animals. Moriah Galvan and Jennifer Vonk (2016) recently tested several cats to measure their capacity to respond to emotional cues coming from their owners and unfamiliar individuals. When the owner displayed a bodily posture or facial expression that signalled happiness, the cats spent more time in contact with them and displayed more positive behaviours than when the cues coming from the owner signalled anger. There was no significant difference in their behaviour when the cues came from the stranger. In a second experiment, the cats were placed inside a carrier and into a room where their owner and an experimenter were engaging in an emotionally charged conversation, with either a positive or a negative valence. The cats did not appear to respond differently to either condition, as there was no significant difference across trials in their proximity to both actors, their latency in exiting the carrier, or their behavioural dispositions. The authors hypothesised that this may have been due to the fact that the actors were not displaying natural emotions, but merely acting, and so certain chemical or olfactory cues that cats purportedly use to discriminate emotions may have been missing.

Horses have also been recently tested for their capacity to discriminate between different emotions in humans. Amy Victoria Smith *et al.* (2016) presented horses with pictures of human faces displaying happy or angry facial expressions, and the spontaneous physiological and behavioural responses of the horses were measured. The subjects preferred to view the angry stimuli with their left eye, which is a sign of right hemispheric bias—a form of lateralisation that is associated with stimuli that individuals perceive as negative. Further, the measures performed on the horses' heart rate showed a faster increase when they were viewing the photographs of angry human faces. The experimenters interpreted these results as providing "the first evidence of horses' abilities to spontaneously discriminate, both behaviourally and physiologically, between positive and negative human facial expressions" (Smith et al. 2016, 4).

With respect to dogs, the last few years have seen a growing amount of evidence regarding their ability to discriminate human emotions. Miho Nagasawa *et al.* (2011) found that dogs can learn to discriminate pictures of humans smiling from pictures of humans with a neutral facial expression and that they can generalise this discrimination to novel human faces. A study performed by Deborah Custance and Jennifer Mayer (2012) found that dogs behaved in a markedly different manner when they were confronted by a human (their owner or a stranger) who was crying, as opposed to one who was merely talking or humming. In the former cases, the dogs oriented towards the human significantly more often, and when they approached them, they did so in a submissive, rather than playful, alert, or calm, manner.

David Buttelman and Michael Tomasello (2013) tested dogs' capacity to interpret human emotional expressions by presenting them with two boxes and a human actor who looked at the contents of each box and displayed, in turn, either a facial expression and vocalisation that signalled happiness or one that signalled disgust. The dogs were then left to choose which box to be fed from, and were found to reliably pick the one that the actor had reacted happily to. No difference in their performance was found when the actor displayed a happy reaction to one box and a neutral one to the other. In a somewhat similar study by Isabella Merola *et al.* (2013), dogs observed their owners reacting in a happy manner towards the contents of one box, and then in a fearful manner towards the content of another box. When given the choice, the dogs preferred to explore the box whose contents had elicited a happy response in their owner. However, when the informant was a stranger, the dogs showed no statistically significant preference for either box.

There is, therefore, a significant, and growing, amount of evidence that points to a capacity, in some mammalian species, to discriminate different affective states in others. And some of these studies specifically suggest the presence in animals of an ability to detect *distress* in others, where this term is understood to encompass different affective states with an extreme negative valence, such as pain, fear, sadness, or anxiety. The evidence gathered to date suggests the presence of this ability in monkeys (Miller, Murphy, and Mirsky 1959; Miller, Banks, and Ogawa 1963; Miller, Caul, and Mirsky 1967), great apes (Berntson et al. 1989; Parr 2001; Slocombe, Townsend, and Zuberbühler 2009), elephants (Plotnik and de Waal 2014), rats (Nakashima et al. 2015), deer (Lingle and Riede 2014), and dogs (Custance and Mayer 2012; Merola et al. 2013).

All of this evidence can, of course, be reinterpreted to fit the behaviour reading hypothesis. In order to pass these tests, it is not necessary that any of these animals be capable of conceptualising distress, happiness, anger, disgust, and so on as *mental* states. They only need to be capable of reading, in an appropriate manner, the behavioural cues that accompany these affective states. This is perhaps more evident in those experiments that simply test the animals' capacity to discriminate one affective state from another, where what the animals are presented with is, precisely, the superficial behavioural cues that correspond to each affective state, such as facial expressions (e.g. Miller, Murphy, and Mirsky 1959), or vocalisations (e.g. Slocombe, Townsend, and Zuberbühler 2009). The animals can simply be discriminating between different behavioural cues. But even the more complex experiments where animals are tested for their capacity to extract information from others' affective states can be reinterpreted along these terms. So, for instance, in the experiment by Merola *et al.* (2013) where dogs chose a box on the basis of their owners' emotional reaction to it, the dogs had perhaps simply learnt to associate their owner's 'happy' facial expressions, bodily postures and vocalisations with positive things, and that is why they chose that box, and not the other. This sort of association would also have permitted the chimpanzees in Parr (2001) to match the pictures of hypodermic needles with those of facial expressions that signify distress, the ones of toys with those of happy facial expressions, and so on.

The most compelling evidence of affective mindreading in a nonhuman species comes from two very recent experiments by Corsin Müller *et al.* (2015) and Albuquerque *et al.* (2016). In the first experiment (Müller et al. 2015), dogs were trained to discriminate pictures showing angry human faces from ones showing happy human faces, and crucially, they were only shown either the upper half or the lower half of the pictures. Once they mastered this task, the dogs were tested for their capacity to extrapolate to the half that had not been used in training. The dogs were found able to categorise pictures of the other half, or of novel faces, according to whether they represented an angry or a happy emotional expression. The subjects could not have been using simple discrimination rules such as 'teeth-showing' or 'teeth-not-showing', as this would not have allowed them to respond adequately to pictures of the different half than the one used in training. According to the authors, "they must have used the emotional expression of the stimuli to solve this task, as this was the only distinguishing feature shared by the training stimuli and the stimuli of all four probe conditions" (Müller et al. 2015, 603).

In the second experiment (Albuquerque et al. 2016), dogs were presented with pairs of pictures that showed the same individual (either human or dog) displaying two different facial expressions (happy/playful and angry/aggressive). While the pictures were being displayed, recordings of vocalisations from the same species with either positive or negative valence were played, and the spontaneous looking behaviour of the experimental subjects was monitored. The dogs were found to direct their gaze significantly more often at the picture that corresponded with the recordings they were hearing, regardless of the stimulus valence and species, although the results were stronger when the pictures and recordings were of dogs. No preference for either picture was found when the stimulus was a neutral vocalisation.

While these two experiments quite strongly suggest that the dogs were classifying the different stimuli using categories that had some relation to their emotional content, the subjects' performance can also be reinterpreted along the lines of the behaviour reading hypothesis. In the first experiment (Müller et al. 2015), the dogs could have learnt to understand the cues 'teeth-showing' and 'relaxed-eyebrows' as both pertaining to the category 'happy-behaviour'—a behavioural abstraction that they would benefit from having, since it would allow them to better predict their owners' behaviour, and know, for instance, when it's time for a walk. Behavioural abstractions of this sort would have allowed them to successfully pass this test without mindreading abilities. The authors themselves hint at this when they say that their results show that dogs "can discriminate between emotional *expressions* in a different species," but that this "does not necessarily mean that the dogs recognized the emotional *content* of the presented stimuli" (Müller et al. 2015, 603, my emphasis). In the case of the experiment by Albuquerque *et al.*, while the authors interpreted the results as proving that dogs possess an "ability to form emotional representations that include more than one sensory modality" (Albuquerque et al. 2016, 4), the subjects could also have passed the test by means of a behavioural abstraction that allowed them to classify both visual and auditory behavioural cues as part of the same class of behaviour, for instance, 'angry-behaviour.'

That this sort of reinterpretation of the data is possible is, luckily, not a problem for my purposes. What is essential for the possession of MME as I have defined it is that creature C be capable of detecting distress *behaviour* in others. To fulfil this condition, C needs to be, at the very least, a behaviour reader, but it is not necessary that she possess mindreading capacities. The moral importance of detecting distress behaviour is, of course, the fact that distress behaviour is a reliable marker of distress, which is, in turn, a morally salient feature of situations (see section 5.3.1), but since we're trying to establish the minimal cognitive requirements for a form of empathy that will qualify as a moral emotion, C is not required to be capable of conceptualising distress as a mental state and attributing it to those displaying distress behaviour. What condition (1) of my definition simply requires is that C be capable of detecting (some of) the different visual, auditory, olfactory, or chemical cues that accompany distress in others, and categorise them in the abstraction 'distress-behaviour,' so that she can distinguish, in a fairly reliable manner, those situations in which individuals around her are displaying distress behaviour, from others in which they are displaying non-distress behaviour.

Recall that behaviour reading implies a certain amount of cognitive activity; an amount that is sufficient for MME to take place. Daniel Povinelli and colleagues, as we saw in section 3.3, insist upon the idea that a behaviour reader can construct abstract behavioural categories and use them to reason about others' behaviour. Since they acknowledge a behaviour reader's capacity to interpret certain behavioural cues ("pursing lips, bristling hair, etc.") in terms of a more abstract behavioural category ("threat display") (Gallagher and Povinelli 2012, 150), it is presumably within the reach of a behaviour reader to interpret certain behavioural cues that are, in fact, markers of distress (e.g. vocalising in a certain way, displaying certain bodily postures, giving off a certain odour, etc.), as pertaining to the more abstract category of 'distress-behaviour.' Therefore, Povinelli's behaviour reader can meet the requirement of being able to identify distress behaviour and represent it as such. Whether this means that C must possess the concept DISTRESS BEHAVIOUR is something that I shall discuss in section 5.2.3. Let us now move on to condition (2) of the definition of MME.

5.2.2. Emotional contagion

Condition (2), which establishes that the detection of distress behaviour by creature C must trigger a process of emotional contagion, constitutes the reason behind my categorising MME as a form of empathy, for it allows me to preserve what is usually seen as the main characteristic of empathic processes. As we saw in chapter 4, it is common for scholars to understand empathy as a process with affective resonance at its core, although this is by no means a unanimously accepted characteristic of empathy (see, for instance: Zahavi and Overgaard 2012; Gallagher 2012). MME is not intended as an addition to the literature on what empathy actually is or what the proper characterisation of this term should be, but simply as a hypothetical example of how moral motivations can obtain in the absence of mindreading capacities. Nevertheless, I have attempted to characterise MME in a way that brings it close to many popular characterisations of empathy. Hence, the form that condition (2) takes.

According to condition (1), as we saw, C has to be capable of detecting distress behaviour in others in a fairly reliable manner. Condition (2), in turn, establishes that C must automatically experience an emotional contagion upon the detection of distress behaviour in another. This means that, upon this detection, C must automatically undergo an affective state that is *isomorphic* to (i.e. of a similar form as) the affective state of the target. While a form of MME involving positive affective states is conceivable (see section 6.2), in the definition I have provided the core affection is a negative one, namely, distress—broadly understood to encompass fear, pain, anxiety, sorrow, and so on. Thus, when C detects distress behaviour in another, she must automatically undergo distress herself, and crucially, C's distress must not be merely coincidental in space and time with the target's, but has to have been triggered by the attended perception of the target's distress behaviour.

Emotional contagion is a phenomenon that has been widely demonstrated in humans (for a review, see: Hatfield, Cacioppo, and Rapson 1994). In the case of animals, the evidence has only just begun to be gathered, but there is already enough to suggest that humans are not the only species to possess this ability. To begin with, we now know that the capacity for automatic motor mimicry, which is often described as one of the mechanisms that can trigger emotional contagion (see, e.g.: Hatfield, Cacioppo, and Rapson 1993; Preston and de Waal 2002), is present in several nonhuman species. Recent experiments have provided evidence for the presence of contagious yawning in some non-human primates, specifically, chimpanzees (Anderson, Myowa-Yamakoshi, and Matsuzawa 2004; Campbell et al. 2009), bonobos (Palagi, Norscia, and Demuru 2014), stumptail macaques (Paukner and Anderson 2006), and gelada baboons (Palagi et al. 2009). Surprisingly, this is an ability that has also been detected in budgerigars (Gallup et al. 2015), as well as wolves (Romero et al. 2014), so it may be present in further mammalian and non-mammalian species. There is also evidence of inter-specific contagious yawning, as chimpanzees (Campbell and de Waal 2014) and dogs (Joly-Mascheroni, Senju, and Shepherd 2008; Romero, Konno, and Hasegawa 2013) have been found capable of 'catching' human yawns. According to another study (Bard 2006), neonatal chimpanzees have the capacity to imitate mouth-opening and tongue-protruding gestures from a human experimenter, a capacity much like the one found in human neonates (Meltzoff and Moore 1977). An observational study further determined that orang-utans engage in involuntary rapid facial mimicry when playing with conspecifics (Davila Ross, Menzler, and Zimmermann 2008). Rapid mimicry during play has been observed as well in gelada baboons (Mancini, Ferrari, and Palagi 2013) and domestic dogs (Palagi, Nicotra, and Cordoni 2015). And lastly, a study found that the pupil size of chimpanzees will tend to mimic that of the observed conspecific (Kret, Tomonaga, and Matsuzawa 2014).

Since evidence of motor mimicry in a species does not necessarily entail the presence of emotional contagion (though both capacities are often linked in the literature—see, e.g. Palagi, Nicotra, and Cordoni 2015), the most relevant evidence for our purposes comes from studies that have specifically focused on affective resonance, rather than motor mimicry. These sorts of studies tend to measure the physiological or behavioural reactions in animals who are witnessing emotional behaviour in other individuals, in search for indications that there is an emotional convergence between both subjects, which may indicate the presence of affective resonance.

In one of these studies, Lisa Parr (2001) measured the skin temperature of chimpanzees who were watching recordings of conspecifics injected with darts and needles, and discovered that their skin temperature decreased during these viewings, something that, in humans at least, is associated with negative emotional arousal. This could be an indication of emotional contagion, although it may also have been an aversive reaction caused by seeing the darts and needles themselves.

Claudia Wascher, Isabella Scheiber, and Kurt Kotrschal (2008) measured the heart rate of free-ranging greylag geese confronted with social events (agonistic interactions between affiliated or non-affiliated conspecifics) and compared the results with their reaction to non-social events, such as vehicles passing by and loud noises. Their heart rate was found to increase significantly more in response to social events than to non-social ones, even though the latter were often more intense (e.g. noisier) than the former. This increase in heart rate was more pronounced when the individuals involved in the agonistic interaction were affiliated with the subjects. While the authors do not commit to an interpretation of these results in terms of emotional contagion, they do point to this as the most plausible explanation (Wascher, Scheiber, and Kotrschal 2008, 1657). They back up this interpretation by referring to the discovery of mirror neurons in birds (Prather et al. 2008). Emotional contagion has also been postulated to be enhanced by social closeness or familiarity (Preston and de Waal 2002), an idea that is consistent with this and with further evidence, as we shall see.

Amanda Jones and Robert Joseph (2006) measured hormone changes in human-dog pairs who had taken part together, and lost, at a competition. The human subjects presented higher testosterone levels after each loss, and this was correlated with higher cortisol levels (a hormone associated with stress) in their dogs. A possible explanation for this effect may be inter-species emotional contagion. In another study, Zsófia Sümegi, Katalin Oláh, and Jószef

Topál (2014) investigated whether they could exert changes on dogs' affective states by manipulating solely the affective state of the owner (giving them a positive or a negative feedback after their completion of a cognitive task). Those dogs that interacted with a stressed owner performed better in a working memory task, which the authors interpreted as a plausible indication of emotional contagion, since stress can improve cognitive performance.

The most compelling evidence of canine emotional contagion, however, comes from a study performed by Min Hooi Yong and Ted Ruffman (2014). They measured the cortisol levels of dogs before and after they listened to a recording of a human infant crying, and compared the results to those obtained in humans subjected to the same stimulus. Both species showed a similar increase of cortisol levels after listening to the crying sounds. This increase was not present in control tests using infant babbling and white noise. The dogs also presented alert and submissive behaviour while listening to the infant crying. The authors interpreted these findings as evidence of an ability for cross-species emotional contagion in dogs.

Some of the most compelling evidence for emotional contagion in a nonhuman species comes from studies on rodents. Dale Langford *et al.* (2006) investigated whether the presence of another conspecific in pain would alter the pain behaviour of mice injected with acetic acid (a painful stimulus). Mice were placed in pairs within separate Plexiglas containers, and then either one of them was injected, or both of them were. The mice were found to display significantly more pain behaviour (what is termed 'hyperalgesia') when their partner had been injected too, provided that they had both previously been cagemates. The pain behaviour also appeared to co-occur in time, something that was observed even when the mice were strangers. By blocking different sensory inputs, the authors found that the determining factor was the visual access to their partner. That is, when the mice could neither hear nor smell their partner, they would still display the hyperalgesia, but this reaction, as well as the co-occurrence of pain behaviour, disappeared when they were separated by an opaque barrier that prevented them from seeing each other. The experimenters further discovered that subjecting each mouse to a different dosage of formalin (a painful stimulus) caused them to influence each other's pain behaviour, provided that they were cagemates. Thus, the mice receiving a lower dosage than their partner showed an increase in their pain behaviour, and the one receiving a higher dosage displayed a decrease. Additionally, when each mouse was subjected to a painful stimulus of a different sensory modality (acetic acid in one case, and a heat stimulus in the other), cagemates also displayed hyperalgesia. The authors interpreted these findings as suggesting that "the pain system is sensitized in a general manner by the observation of pain in a familiar" (Langford et al. 2006, 1969).

Ewelina Knapska *et al.* (2006) studied the contagion of fear in rats. Across the study, the subjects were kept in pairs, and then one of them (the demonstrator) was extracted from the cage and subjected to a stimulus that elicited fear (a different cage that delivered an electric shock to its feet). The demonstrator was then returned to the cage in order to measure the behavioural reaction of its cagemate (the observer) to the former's conditioned fear. The authors determined there had been "a transfer of emotional state between subjects" (Knapska et al. 2006, 3859), since the observers presented an activation of their amygdalae (associated with fear) to levels similar as that of the demonstrator, as well as behavioural patterns characteristic of fear and an intense exploration of the demonstrator, none of which was present in control conditions.

Daejong Jeon *et al.* (2010) found that mice who observed conspecifics receiving repetitive foot electroshocks displayed freezing behaviour, a fear response that could be a sign that emotional contagion was taking place. This response took place even when the observers had never experienced foot shocks, and was more intense when both individuals were socially related. A similar study performed on rats (Atsak et al. 2011) also found that witnessing a conspecific in pain induced freezing behaviour, although these results were only obtained when the observers themselves had previously been subjected to foot shocks.

James Burkett *et al.* (2016), in a study we saw in section 2.1.1.3, investigated whether prairie voles (a type of rodent) would respond with consolation to a conspecific in distress. In order to do this, they separated pairs of voles who were cagemates and subjected one of them (the demonstrator) to tones followed by electric shocks, in order to induce Pavlovian fear conditioning. The demonstrator's cagemate (the observer) was then reintroduced into the room, and her behaviour observed. Given that they found an increase in allogrooming consistent with the hypothesis that the observer was engaging in consolation behaviour towards the demonstrator, the researchers set out to investigate whether an empathy mechanism could be at the basis of this behaviour. They found that "prairie vole observers showed behavioral responses consistent with emotional contagion by mimicking the anxiety- and fear-related behaviors of stressed demonstrators" (Burkett et al. 2016, 376). In addition, when the demonstrator freezed as a result of hearing the conditioned stimulus (the tone that was usually followed by an electric shock), the observer displayed freezing behaviour as well, which is also consistent with emotional contagion of fear.

There is, lastly, some evidence concerning emotional contagion in farm animals, which has attracted some attention due to the consequences that it would have for the welfare of these species if they were to be emotionally affected by the distress of others—a stimulus that they are often exposed to in intensive farming conditions. While studies performed on sheep (Anil et al. 1996; Colditz, Paull, and Lee 2012) have thus far failed to confirm the presence of emotional contagion in this species, suggestive results have been obtained in studies on chickens and pigs. Joanne Edgar *et al.* (2011) subjected hens to a mildly aversive stimulus (air puff sprayed into their cage) and allowed them to witness their chicks subjected to the same stimulus. They discovered that the hens would undergo similar physiological and behavioural reactions (associated with mild distress) in both conditions, which they interpreted as showing that adult female chickens possess "the ability to be affected by, and share, the emotional state of another" (Edgar et al. 2011, 238). However, this ability may be limited to the mother-offspring bond, since another study (Edgar et al. 2012) failed to replicate these results when the targets were familiar adult conspecifics. In the case of pigs, Inonge Reimert *et al.* (2013; 2015) found that members of this species will display behaviours associated with positive emotions (e.g. play, bark, tail movements) or negative ones (e.g. ears back, tail down, urinating, defecating) depending on whether they have been paired with a conspecific who has undergone a positive or a negative treatment.

While the evidence is not extensive, it is enough to indicate that emotional contagion may be present in several nonhuman species. Additionally, it is important to bear in mind that the majority of these studies are extremely recent, and that this is an issue that has only just begun to gather scientific attention. It is thus expected that more relevant evidence will be obtained in upcoming years. For our purposes, however, it suffices. I wanted to argue for the empirical plausibility of condition (2), according to which, creature C has to experience an emotional contagion upon the detection of distress behaviour in another. The evidence I have reviewed suggests that this is an ability that is present, at the very least, in chimpanzees (Parr 2001), greylag geese (Wascher, Scheiber, and Kotrschal 2008), dogs (Yong and Ruffman 2014), mice (Langford et al. 2006; Jeon et al. 2010), rats (Knapska et al. 2006; Atsak et al. 2011), prairie voles (Burkett et al. 2016), chickens (Edgar et al. 2011), and pigs (Reimert et al. 2013; 2015). This is more than enough to consider it an empirically plausible condition.

5.2.3. Intentional emotional contagion

Condition (3) establishes that the emotional contagion that creature C undergoes must give rise to an emotion that has the target's distress behaviour as its intentional object. This means that the result of the emotional contagion has to be that C is distressed *that* the other is displaying distress behaviour; this is what C must be distressed *about.*

This is the crucial characteristic that distinguishes my account of empathy from other accounts in the literature, and it is here where we find the key to a minimal conception of empathy as a moral emotion. This condition is novel, but has been inspired by considerations in Rowlands (2012b). In putting it forward, I am following this author's analysis of moral emotions (which we already saw in section 2.3.2). According to this analysis, moral emotions must contain (but not necessarily be reducible to) two different types of content: factual content and evaluative content. The factual content is the state of affairs that the emotion is directed at. In the case of C, this factual content can be specified by use of the proposition "This creature is displaying distress behaviour." Such factual content is what C entertains emotionally as distressing. With respect to the evaluative content ("This creature's distress is bad"), I have already sketched how Rowlands' framework allows this sort of content to be present in an individual's emotion without requiring a capacity to explicitly entertain it. This is an issue that I will return to in section 5.3.2, so, for now, let us focus on the factual content of MME.

For MME to take place, then, the distress that is triggered by C's emotional contagion must include the factual content "This creature is displaying distress behaviour." This is the same as saying that the presence of a creature displaying distress behaviour in her immediate environment has to be what C is distressed *about*. What other things could C be distressed about as a result of her emotional contagion? If C could mindread, then the factual content of her distress could potentially be the proposition "This creature is distressed." We are assuming, however, that no animal can mindread, so I shall leave this possibility aside. We can still distinguish (at least) three further alternatives:

1. C's emotional contagion could give rise to a simple mood, in which case, her affective state would have no intentional object. In such a scenario, C would have become automatically distressed as a consequence of her detection of the target's distress behaviour, but her resulting distress would be about nothing in particular—a simple affective state with no cognitive content. This is, presumably, what happens when a newborn baby starts crying upon hearing other infants crying (Sagi and Hoffman 1976). If this were indeed the case, it would mean that there is nothing in the world that the baby is distressed about; her affective state is simply the result of a purely physiological mechanism.
2. The intentional object of C's distress could, alternatively, be whatever is causing, or is

thought to be causing, the distress of the target. The chimpanzees in Parr (2001), for instance, might have become distressed as a result of an emotional contagion triggered by watching their conspecific's reaction to being injected, but the resulting emotion perhaps had, as its intentional object, the hypodermic needles themselves, and not the target's distress behaviour. The factual content of the chimpanzee's distress would be the proposition "There is a hypodermic needle," or perhaps (depending on how the chimpanzee represents the object in question) "There is that pointy thing that hurts." This sort of intentional object may obtain even if the cause of the target's distress isn't perceptually available. To see this, imagine a flock of birds that are foraging in a field. One of them is startled and takes off; the rest, immediately, follow. The latter may be due to an emotional contagion that has given rise to a form of distress directed at the possibility of there being a predator around. The factual content could then be specified by the proposition "There might be a predator around." In this case, the cause of the target's distress is not perceptually available, and yet it may still be the intentional object of the birds' fear.

3. Lastly, there is the possibility that the emotional contagion triggers a form of distress that has C's distress itself as its intentional object. This might be the case if C becomes so distressed as a result of the process of emotional contagion that her own distress becomes her sole concern. This comes close to what psychologists often call 'personal distress,' which has been defined as "a *self-focused*, aversive affective reaction to the apprehension of another's emotion, associated with the desire to alleviate one's own, but not the other's distress" (Eisenberg and Eggum 2009, 72, my emphasis). Personal distress can also be the result of an incapacity to distinguish the self from others (see, e.g.: Jean Decety and Lamm 2009, 199). If C doesn't possess a clear self-other distinction, then she might interpret the target's distress, which she catches as a result of her emotional contagion, as her own. Her resulting distress might then have as its factual content the proposition "I am distressed," or "This is unpleasant," which would

presumably trigger a search for an alleviation of her own distress.

These three alternatives are all possible results of a process of emotional contagion. Whether an individual instance of emotional contagion triggers any of these three possibilities is likely dependent upon different factors, such as the individual's hardwired dispositions, past learning experiences, or other intervening cognitive mechanisms. My interest is not in arguing that emotional contagion will trigger a form of distress with a specific intentional object under such and such circumstances. The way I have defined MME (which, remember, is just a set of *sufficient* conditions for a minimal form of empathy as a moral emotion) requires C's emotional contagion to trigger a form of distress with the target's distress behaviour as its intentional object. While I do believe that this (coupled with the other conditions I have put forward) would be enough for MME to qualify as a moral emotion, I take no stand on what empirical conditions are necessary for emotional contagion to effectively trigger an emotion of these characteristics.

The alternative forms of emotional contagion we have considered have their focus on the individual's own well-being, and, for this reason, are likely to trigger entirely self-centred behaviours. There is, therefore, no *prima facie* reason to suppose that they are moral in character. I will thus proceed by bracketing them and reserve the label *intentional emotional contagion* to refer to one that triggers a form of distress directed *specifically* at the state of affairs that is the target displaying distress behaviour. What evidence do we have, then, of the empirical plausibility of intentional emotional contagion? Unfortunately, this is a novel notion, and thus has not been subjected to direct empirical investigation. It is therefore not possible to offer a selection of empirical examples that apply to this condition, as has been done in the previous sections. However, it seems reasonable to grant that intentional emotional contagion as I have defined it is a plausible notion, given its strong resemblance to aversive arousal, which, as we saw in section 4.3, is widely proposed by scholars as a possible explanation for the performance of animals in various morality and prosociality tests. As the reader may recall, aversive arousal is understood as a reaction in which the target's distress behaviour is perceived as an aversive stimulus. In the case of intentional emotional contagion, the content of C's distress is the other's distress behaviour. This amounts to experiencing the target's distress behaviour as something distressing, which comes remarkably close to experiencing it as an aversive stimulus, insofar as something that is distressing will, by definition, be unpleasant, and whatever is unpleasant, in turn, characteristically elicits in the subject a strong dislike, or *aversion*. If aversive arousal is generally regarded as a plausible explanation of the evidence, I see no reason why intentional emotional contagion should not be viewed, as well, as an empirically plausible mechanism.

Some scepticism may remain in the reader as a result of my characterisation of intentional emotional contagion as a process that generates an emotion with propositional content. What do I mean when I say that C undergoes a form of distress that has the proposition "This creature is displaying distress behaviour" as its factual content? I merely mean that C has the capacity to detect, in a fairly reliable manner, instances of distress behaviour in others, which she represents using a single abstract category—the category 'distress-behaviour'—, and that when she undergoes such detection, her emotional contagion is triggered, and this yields a form of distress that is directed at the distress behaviour of the target *represented as distress behaviour*. The purpose of bringing up how C must represent the target's distress behaviour is to characterise C as a behaviour reader. If C were a mindreader, then she could, for instance, undergo a form of distress directed at the distress behaviour of the target *represented as a cue that signals distress*. Or she might be able to *see*, in the epistemic sense of the term (see section 4.1.3), *that* the target is distressed, because she possesses the concept DISTRESS. But because I want to argue for the possibility of morality without mindreading, it is important to characterise C as a behaviour reader, which means she must represent the other's distress behaviour *as distress behaviour* (see section 5.3.1 below for further considerations on the importance of this requirement).

That animals capable of possessing emotions can undergo emotions that are directed at particular objects or states of affairs in the world should be fairly uncontroversial. At least, anyone who has seen a dog react to a vacuum cleaner should have no doubt that members of this species can be afraid *of* specific objects in the world. Of course, these sorts of anecdotal observations could be attributed to an incurable human tendency to explain everything in folk psychological terms. However, it seems fairly obvious that, once a being possesses emotions (which, according to the available evidence, is probably the case for at least all mammalian species; see: Panksepp 2004), there are huge evolutionary advantages to being capable of feeling emotions that are directed at particular objects in the world. The gazelle who fears the lion will undoubtedly have more chances of survival than the one who simply experiences a non-intentional form of fear. If there is any cause for controversy, this is likely to be the idea that animals can possess emotions that are directed at objects *represented using an abstract category*. While it is fairly uncontroversial to say that the gazelle fears the lion, it is more problematic to say that she fears an individual that she represents as belonging to the abstract category of 'predator,' or, to put it differently, that she can undergo a form of fear with factual content that we can specify by use of the proposition "There is a predator."

MME requires that C be capable of detecting, in a fairly reliable manner, the different forms that distress behaviour may take, and categorise them using one single abstraction. This can be interpreted as the requirement that C possess the concept DISTRESS BEHAVIOUR—this is, in fact, the way that Daniel Povinelli and colleagues understand the behavioural abstractions that, according to them, behaviour readers are capable of forming. They assert, for instance, that "[c]himpanzees undoubtedly form *concepts* related to the statistical regularities in behavior" (Povinelli and Vonk 2003, 157, my emphasis). Condition (3), then, could be interpreted along these lines. Since my primary aim is to argue for the possibility of moral emotions in behaviour readers, it could perhaps be enough to assert that Povinelli *et al.* would presumably grant me this requirement. However, due to the fact that the attribution of concepts to animals is still a hotly debated issue in animal cognition, some readers may feel that more needs to be said. Let us thus go into this topic in a little more detail.

The first point that needs to be made is that many of the arguments that have been provided against the possibility of attributing concepts to animals are not ontological in nature, but epistemological. What is often argued is not that animals don't possess concepts, but that it is impossible for humans to know whether they do in fact possess concepts, or the form that purported concepts take in their minds. For instance, in one of the most famous texts on this topic, Donald Davidson asserts:

> [C]an the dog believe of an object that it is a tree? This would seem impossible unless we suppose the dog has many general beliefs about trees: that they are growing things, that they need soil and water, that they have leaves or needles, that they burn. There is no fixed list of things someone with the concept of a tree must believe, but without many general beliefs, there would be no reason to identify a belief as a belief about a tree … If we really can intelligibly ascribe single beliefs to a dog, we must be able to imagine how we would decide whether the dog has many other beliefs of the kind necessary for making sense of the first. It seems to me that no matter where we start, we very soon come to beliefs such that we have no idea at all how to tell whether a dog has them, and yet such that without them, our confident first attribution looks shaky. (Davidson 1982, 320-1)

The claim is, then, that we cannot engage in a confident attribution of a belief to an animal unless we can be sure that she possesses the concepts it is made up of. At the same time, we cannot be sure that the animal possesses these concepts unless we know that she has a certain number of relevant beliefs about the objects that fall under their scope. These are beliefs that refer to properties that objects have in virtue of the fact that they are instances of the concepts themselves. So, in the case of the concept we're interested in, DISTRESS BEHAVIOUR, Davidson might have said that C cannot be attributed this concept unless we're sure that she possesses a set of different beliefs about distress behaviour, such as the belief that it is an unlearned physiological reaction, that it is usually caused by aversive stimuli, that it is present in many different species, that it encompasses a number of different behaviours, and so on. Attributing these beliefs to C would, in turn, require attributing the concepts they are composed of, which, once again, requires attributing further beliefs, and so on. Inevitably, according to Davidson, we will reach certain beliefs whose possession on behalf of C we have "no idea at all" how to ascertain, but in the absence of which the initial attribution seems unwarranted.

There are several steps in Davidson's argument that could be questioned. For our purposes, however, the most important flaw in it is that it cannot be used to argue that animals don't possess concepts. At most, it could be used to argue that we can never *know* if animals do in fact possess concepts. The claim Davidson is ultimately making is epistemological, and not ontological.[5] The argument famously put forward by Stephen Stich (1979) falls prey to an identical objection. The essence of what he defends is condensed in the following quote:

> Our difficulty in specifying the contents of animals' beliefs derives not from an ignorance of animal psychology but rather from a basic feature of the way we go about assigning content to a subject's beliefs: *We are comfortable in attributing to a subject a belief with a specific content only if we can assume the subject to have a broad network of related beliefs that is largely isomorphic with our own.* When a subject does not share a very substantial part of our own network of beliefs in a given area we are no longer capable of attributing content to his beliefs in this area. (Stich 1979, 22, his emphasis)

[5] In his other most famous argument on the topic, Davidson does make an ontological claim. He argues that since animals lack linguistic capacities, they cannot possess the concept of belief, and given that the latter is a necessary requirement for the possession of beliefs, this means that animals cannot have beliefs (Davidson 1982, 324ff.). This argument has been repeatedly questioned on a variety of different grounds. In fact, Davidson himself didn't take it to be very persuasive (Davidson 1982, 324). For these reasons, and to avoid yet another detour in this chapter, I shall not address it here. Examples of counter-arguments can be found in texts such as Searle (1994), Glock (2000), Andrews (2002), Newen and Bartels (2007), Rowlands (2009, chap. 7), Diéguez (2011), Rowlands and Monsó (forthcoming).

Nonhuman animals, according to Stich, can be confidently classified as "subjects whose total collection of beliefs [differs] significantly from our own," which means that nothing we could do would ever allow us to specify the content of their beliefs, i.e., the concepts that constitute them (Stich 1979, 24). If animals do indeed have beliefs, their content is non-specifiable. Stich explicitly refuses to take a stand on whether the lack of specifiable content means no content at all (Stich 1979, 27), so his argument, as Davidson's, is merely epistemological. Similarly, Nick Chater and Cecilia Heyes have defended the idea that humans lack any viable "account of what cognitive feat possessing a concept amounts to for nonlinguistic agents," which means we "cannot assess whether this or that animal possesses concepts or not, still less ascertain the content of its putative concepts" (Chater and Heyes 1994, 210). Their claim is not that animals lack concepts, but that "content ascription is likely to be rough and ready for the foreseeable future, and should be recognized as such" (Chater and Heyes 1994, 239).

A major concern, therefore, when it comes to talking about concepts in animals is the empirical tractability of the problem. In particular, there are concerns about the difficulties involved in answering two questions: (a) *whether* members of any nonhuman species indeed possess concepts and, if so, (b) *which* concepts they possess. Colin Allen (1999) attempted to solve problem (a) by providing a framework for determining whether an animal possesses a certain concept. He states:

> An organism O may reasonably be attributed a concept of X (e.g., TREE) whenever:
>
> (i) O systematically discriminates some Xs from some non-Xs; and
>
> (ii) O is capable of detecting some of its own discrimination errors between Xs and non-Xs; and
>
> (iii) O is capable of learning to better discriminate Xs from non-Xs as a consequence of its capacity (ii). (Allen 1999, 36-7)

These conditions are not intended as a philosophical analysis of what it means for an animal to possess a concept, nor does Allen understand them as necessary, or even sufficient, conditions. His purpose in presenting these three clauses is to provide a set of empirical conditions under which it would be "reasonable to attribute a concept to an animal" (Allen 1999, 37). Fulfilment of these three clauses would provide good evidence that the animal possesses "a representation system that compares the perceptual content with an independent representation of what the perception is supposed to represent, i.e, a concept" (Allen 1999, 39). It may nevertheless be the case that an animal fulfils none, just one, or only two of these clauses and still possesses concepts.

Allen's framework provides us with an empirically tractable way of determining whether an animal possesses concepts. He thus offers a solution to problem (a). However, it could be argued that problem (b), the problem of exactly which concepts an animal possesses, is not entirely solved by Allen's framework. Thus, suppose we find that C (i) systematically discriminates creatures that are displaying distress behaviour from creatures that are not displaying distress behaviour, and furthermore, that C (ii) can detect errors in her discrimination, and (iii) is capable of learning to better discriminate distress-behaviour-situations from non-distress-behaviour situations. We would then have fairly good evidence that C possesses the concept DISTRESS BEHAVIOUR. However, since all instances of distress that are detectable are, by definition, expressed via behavioural cues, it could be the case that C is using the concept DISTRESS rather than DISTRESS BEHAVIOUR. Could Allen's framework provide us with a way of distinguishing both possibilities? If C were found unable to discriminate instances of feigned distress from instances of real distress, it could be argued that we have evidence that C is using the concept DISTRESS BEHAVIOUR rather than DISTRESS. And yet, it seems perfectly conceivable that C is using the concept DISTRESS and is simply unable to tell the difference between feigned and real distress. Thus, while Allen's framework provides us with a way of determining *whether* an animal possesses concepts, it cannot always tell us *which* concepts an animal possesses.[6] Having said that, we should bear in mind that the example I have just used is especially thorny—as we saw extensively in chapter 3, finding a way of empirically distinguishing the use of behaviour concepts from the use of mental state concepts on behalf of animals is proving to be one of the biggest challenges in contemporary comparative psychology. Luckily for us, we need not worry about this problem, since my framework is made to accommodate the least intellectually demanding explanation of animals' performance in cases like this.

Albert Newen and Andreas Bartels (2007) have further argued that Allen's clauses, if taken as a set of necessary and sufficient conditions, would provide too stringent a requirement for the possession of concepts. They consider that "the capacity of correcting one's own errors is not necessary for having concepts even if it may be sufficient" (Newen and Bartels 2007, 291). This is an important point for our purposes. As we saw in section 5.1.4, the morality of MME can obtain even when C mistakes non-distress behaviour for distress behaviour; even if

[6] Mark Rowlands attempts to solve, or rather, *dissolve* this problem by use of the concept of tracking (Rowlands 2012b, chap. 2). The basic idea is that, even if we can't know which concepts an animal possesses, we can still describe her behaviour by means of the attribution of propositional attitudes, so long as the concepts we use in these attributions *track* the concepts that the animal possesses. I have decided not to make use of this idea because it may be a source of confusion for the reader. The problem is that the notion of tracking, in this context (i.e., in reference to factual content), is used in a different sense than when discussing the evaluative contents of animals' (moral) emotions. When discussing factual content, the notion of tracking is used to express the idea of there being a "reliable asymmetric connection" between the concepts *we use to describe* the animal's behaviour and the concepts *the animal, in fact, possesses* (Rowlands 2012b, 58). The animal is thus assumed *capable of entertaining factual content*—we just may be incapable of knowing exactly the form it takes. The notion of tracking here serves as a tool to sidestep this worry. As we've seen, the notion of tracking is also used as a way of specifying the evaluative content of an animal's emotion, which is a type of content *that the animal cannot entertain*. There is, thus, a very clear difference between both uses of the notion of tracking. In the case of factual content, the notion of tracking only works if the animal is capable of entertaining this kind of content to begin with. In the case of evaluative content, the animal need not be capable of entertaining content of this sort. Due to the fact that the notion of emotions tracking evaluative content plays a crucial role in my argument, it is very important to avoid confusions on this point. This is why I won't utilise the solution offered by Rowlands to deal with the problem of factual content.

C’s emotion is, in Rowlands’ terminology, *misplaced* (see also section 5.3.2), so we need not consider the capacity to correct one’s own errors in discrimination as a necessary requirement for MME to take place. This could perhaps be taken as a reason for renouncing the somewhat problematic idea that MME requires the possession of concepts. However, as we saw, it is very important that we characterise C in a way that explicitly excludes her from being considered a mindreader, and thus it becomes crucial for C to predict, explain, and interpret the behaviour of others solely in behavioural terms. At the same time, it is crucial that C’s distress be not only *triggered by* but also *directed at* the target’s distress behaviour. Therefore, even if C cannot correct her own errors in discrimination, it is important that she possess, if not the concept DISTRESS BEHAVIOUR, something that can perform a similar epistemic role when evaluating situations.

A promising way of solving this impasse is to use the framework offered by Newen and Bartels themselves. They understand concepts to be types of mental representations that are individuated by the epistemic capacities they grant the individual who possesses them (Newen and Bartels 2007, 284). In particular, concepts allow the individual to “represent the same property while dealing with different objects and … represent different properties while having only one and the same visual input of an object” (Newen and Bartels 2007, 293). With this in mind, they provide a set of necessary and sufficient conditions for concept possession. It is important to note that these authors consider that concept possession is “not an either-or question, but something that develops gradually” (Newen and Bartels 2007, 294). Their conditions thus provide us with a good opportunity for reflecting on the epistemic capacities that C should have and determining whether MME does indeed require concepts or merely something that is almost-but-not-quite a concept. Let us thus have a look at the conditions they put forward:

> A cognitive system can have the concept RED only:
>
> C_1. If the system has a stable representation of the property of being red while dealing

> with very different red objects, and
>
> C_2. If the system can not only represent the property of being red but also some other properties of the same object.
>
> …
>
> C_3. If we have relative stimulus independence such that it depends on some additional mechanism—which detects and weighs stimuli other than the key stimulus of redness—to determine that the system focuses on redness while perceiving a red square, in contrast to some other property; … and
>
> C_4. If the property of being red is represented in a minimal semantic net: the property *red* is represented as an instance of the dimension color—not as an instance of an incorrect dimension (e.g., danger). (Newen and Bartels 2007, 296)

C_1 is certainly a condition that we most certainly want C to fulfil. C should have a stable representation of the property of displaying distress behaviour so that she can identify and reidentify different instances of this property. And, of course, it is not enough for her to have this representation when dealing only with one individual, e.g., her son. Otherwise, she would not have a representation of the property *displaying-distress-behaviour*, but only a capacity to distinguish *son-displaying-distress-behaviour* situations. She has to be fairly good (although certainly not infallible) at evaluating different individuals using this same representation. This should be an uncontroversial requirement, since we've seen that animals from different species are quite competent at detecting distress behaviour, a capacity that some even appear to be able to exercise in cross-species contexts (e.g. Lingle and Riede 2014; Custance and Mayer 2012).

C_2, the requirement that C be capable of representing further properties of the same object, is also important. Firstly, because it serves to add empirical plausibility to MME. While our primary interest is in attributing to C a capacity to represent the property *displaying-distress-behaviour*, it would make a rather implausible characterisation of C if we were to think of her as a creature who has developed a capacity to attribute one sole property to other creatures. If C is to be a behaviour reader, and not simply a detector of distress behaviour, she needs to be capable of discriminating further behavioural properties in others. Furthermore, as Newen and Bartels point out, if C could only discriminate the property *displaying-distress-behaviour* in others, but not other properties—such as *displaying-threat-behaviour*, or *displaying-playful-behaviour*—then it would not make much sense to say that she is classifying the target as having the property *displaying-distress-behaviour*. Instead, we would have to say that she can merely distinguish distress-behaviour situations from non-distress-behaviour situations (Newen and Bartels 2007, 296). But we surely want C to understand *displaying-distress-behaviour* as a property of the target. This will diminish the probability of C reacting with mere personal distress to her emotional contagion, and will also help to justify the morality of her reaction (see section 5.3).

With respect to C_3, there is a clear sense in which we want C's stable representation to be stimulus independent. We want C to be capable of not simply generating stimulus generalisations, but also of forming abstract categories of behaviour. Her stable representation should thus involve the formation of a class (Newen and Bartels 2007, 295). This is helpful insofar as it allows us to say that C must not only recognise one type of distress behaviour (e.g. distress vocalisations), but several different types (e.g. bodily postures, facial expressions, chemical signals…), all of which she classifies using the same stable representation. This will be very important when it comes to arguing that C has a reliable sensitivity to a morally relevant feature of situations (see section 5.3.1). Additionally, and in contrast to the case of the concept RED, this stimulus independence is relatively easy to test for in the case that concerns us. To get quite good evidence of this stimulus independence, we would only have to compare C's behavioural and physiological reactions to different forms of distress behaviour and determine the extent to which they coincide.

The last condition, C_4, is the trickiest. Do we want C to represent the property of displaying distress behaviour in a minimal semantic net? Ideally, yes. We would want C to understand distress behaviour as a type of behaviour, as opposed to a mental state. But since we are assuming that C can't mindread, perhaps such a requirement would be superfluous, since there is no way that C could understand the target's distress behaviour as a mental state. In other words, since she cannot *epistemically see* the target's distress in her distress behaviour, and cannot *infer* its presence either because *ex hypothesi* she lacks mental state concepts, then it may be unnecessary to require that she classify distress behaviour as a type of behaviour. Ideally, C should of course be able to understand distress behaviour as an instance of the dimension 'behaviour,' which would allow her to have a better understanding of the intensional features of this concept (Newen and Bartels 2007, 297). This ability would also allow C to acknowledge that properties from the dimension 'behaviour' are different from other properties present in individuals, such as the ones pertaining to the dimension 'physical characteristics.' If C fulfilled C_4, then she might be capable of sorting properties such as *displaying-distress-behaviour* and *displaying-threat-behaviour* into one group, and properties such as *female* and *male* into a different group. While this would certainly make C more intelligent, and perhaps more skilled at social interactions, I wouldn't go as far as requiring it for MME. So long as the other three conditions are fulfilled, I see no *prima facie* reason why we need C_4.

Having analysed these four conditions, it seems clear that we definitely want C to fulfil C_1, C_2, and C_3. Since we're trying to construct a *minimal* account of empathy as a moral motivation, we should let go of any requirements that aren't essential. I thus think that we should relinquish C_4. If we were to follow Newen and Bartels, this would mean that the abstract category of behaviour that C constructs is *not* a concept, for these authors present these four conditions as both necessary and sufficient for concept possession. I won't take a stand on whether a stable representation that fulfils only the first three conditions indeed counts as a concept or not, as it is, in my view, merely a matter of stipulation. What is important is that it remains clear what we are asking of C, namely: that she can (1) identify and reidentify instances of distress behaviour, (2) represent further properties of objects, and (3) classify different forms of distress behaviour using a single abstraction. Whether or not this counts as concept possession is irrelevant for present purposes.

5.2.4. A MOTIVATION TO ENGAGE IN AFFILIATIVE BEHAVIOUR

So far, we have gone over the first three conditions that make up MME. We now know that C must be capable of detecting distress behaviour in others in a fairly reliable manner, that this detection must trigger an emotional contagion, and that the emotional contagion has to result in a form of distress that has the target's distress behaviour as its intentional object. The final condition, condition (4), requires there to be a certain behavioural urge built into C's distress at the target's distress behaviour. I described it in the definition of MME as an urge to engage in other-directed affiliative behaviour. By other-directed affiliative behaviour I mean behaviour that involves engaging with the target, can be described as 'friendly,' and would, under normal circumstances, contribute to an alleviation of the target's distress (Kikusui et al. 2001; Fraser, Stahl, and Aureli 2008; Clay and de Waal 2013; Smith and Wang 2014; Burkett et al. 2016). These are behaviours such as touching, stroking, grooming, nuzzling, or hugging.

I have decided to focus on these kinds of behaviours because they are the simplest 'comforting' behaviours—pertaining, perhaps, to what Shaun Nichols would call the "core cases of altruistic behavior" (Nichols 2001, 427)—, and thus presumably the least cognitively demanding. In focusing on this specific type of behaviour, I am leaving aside reactions that entail adopting a different strategy to eliminate the target's distress behaviour, such as extinguishing its cause. An example of this would be the performance of the rhesus monkeys who refrained from pulling the chain that delivered food to them upon seeing that it caused a conspecific to receive an electric shock (Masserman, Wechkin, and Terris 1964; Wechkin, Masserman, and Terris 1964). Something similar to MME could also motivate behaviours of this sort, but they seem to require the intervention of an inference from the target's distress behaviour to its cause, and so are presumably more intellectually demanding than simply reacting with affiliative behaviour.

An obvious question emerges at this point: if C lacks all capacity to understand that the target is distressed, then how is it possible for C to experience an urge to engage in affiliative behaviour as a result of her emotional contagion? Let's go back to the Higgins-Jane example we used in section 5.1 to discuss this problem, now explicitly thinking of Higgins as a nonhuman animal, such as a dog, and of Jane as his human companion. Suppose, then, that Jane is crying, and that Higgins' emotional contagion kicks in upon his detection of Jane's crying (a form of distress behaviour). As we saw in the previous section, the emotional contagion that Higgins undergoes must be of a special sort, namely, what I have termed *intentional* emotional contagion. This means that it must result in a form of distress that has Jane's distress behaviour as its intentional object. As a result of his emotional contagion, then, Higgins is distressed *that* Jane is displaying distress behaviour. The last thing we need for Higgins to be a moral subject is for this intentional emotional contagion to become a motivation to behave in a comforting manner towards Jane. But how could this possibly occur? What could move Higgins to comfort Jane, given that he can't understand that she is distressed; that she *needs* comforting?

It is very important to provide an account of the way in which the urge to engage in affiliative behaviour that is built into Higgins' distress might have *got there* in the first place. If no such account were provided, MME would inevitable lose empirical plausibility. More importantly, we would be failing to properly address one of the strongest intuitions behind the common assumption that morality requires mindreading, namely, the idea that without mindreading, moral behaviour simply *wouldn't occur*. If we had never developed mindreading skills, a sceptic might say, then we would each go about our own business without caring for each other because we would have no *reason* to care. This sort of intuition is possibly at the basis of Nichols' claim that we need at least a minimal capacity for mindreading in order for altruistic motivation to take place (Nichols 2001; see also section 5.1.3). To tackle this intuition, and the objection it can potentially give rise to, we must explain how Higgins' intentional emotional contagion can motivate affiliative behaviour.

The lack of mindreading capacities that, we are assuming, Higgins suffers from precludes him from behaving on the basis of any practical reasoning that takes as its input considerations about the mental states of others. This means that Higgins cannot engage in affiliative behaviour towards Jane as a result of a practical reasoning such as:

P_1: Jane is distressed.

P_2: Jane's distress is a bad thing.

P_3: If I comfort Jane, that will alleviate her distress.

C: I will comfort Jane.

Such a practical reasoning is certainly beyond Higgins' reach. Entertaining any of the three premises (P_1, P_2, and P_3) would require a capacity to attribute mental states to Jane. Entertaining P_2 would further involve the ability to evaluate those mental states in light of moral concepts, and grasping P_3 would require an understanding of how one's behaviour can have an impact on other individuals' mental lives. Since Higgins lacks mindreading capacities and moral concepts, he can't entertain any of these three premises, and so can't engage in a practical reasoning of this sort to reach the conclusion that he should comfort Jane.

So far, Rowlands' (2012b) framework has allowed us to translate P_1 and P_2 into something that is within Higgins' reach. Higgins can't understand that Jane is distressed, but he can (presumably) detect her distress behaviour and represent it as such. He can't understand that Jane's distress is a bad thing, but his intentional emotional contagion renders him distressed at Jane's distress behaviour, thus making him experience the latter as something distressing. For MME to effectively become a moral motivation, we need to find a way of translating P_3, so that Higgins goes from being distressed at Jane's distressed behaviour to being motivated *by* this distress to comfort Jane.

A promising starting point in our search for the link between intentional emotional contagion and affiliative behaviour is to have a look at Rowlands' characterisation of the notion of 'moral module.' As we saw in section 2.3.2, Rowlands considers that the emotional reactions of a minimal moral subject to morally relevant features of situations must not be merely accidental or contingent, but rather, grounded in the operations of a reliable mechanism (see, e.g. Rowlands 2012b, 230). This reliable mechanism is what he refers to by use of the term 'moral module' (where "module" is not meant in a psychologically realistic sense—see: Rowlands 2012b, 146). Throughout his 2012b book, we can find several characterisations of this 'moral module,' which tend to present it as a mechanism that triggers a certain emotion when certain features are perceptually detected in the environment, more precisely, as a mechanism that "connects perceptions of the morally salient features of a situation with appropriate emotional responses in a reliable way" (Rowlands 2012b, 146). On a couple of occasions, however, this reliable mechanism is presented as providing the link between the perception of certain features, a certain emotional reaction, *and* a certain resulting behaviour. Thus, we can find the assertion that the 'moral module' is what "links perception, emotion, and action" (Rowlands 2012b, 152), and that the moral subject's "actions, emotions, and the morally salient features of the situations in which [she] finds [herself] are all connected by way of the reliable operations" of this mechanism (Rowlands 2012b, 167).

This ambiguity between a characterisation of the 'moral module' as a mechanism that links perception and emotion, and one that links perception, emotion, and action is, quite probably, only apparent. This is so because Rowlands is presumably thinking of the motivation to behave in a certain way as an important component of emotions themselves. This interpretation seems in line with his claim that the deliverances of a 'moral module' would "take the form of emotions or sentiments," which in turn "motivate [one] to think and act in ways that are (morally) appropriate to the exigencies of the situation" (Rowlands 2012b, 153). Thus, when he characterises the 'moral module' as a mechanism that links perception and emotion, the additional behavioural component is implicit in the characterisation insofar as emotions are understood as motivations to behave in a certain way. Even though the ambiguity in Rowlands' argumentation may only be apparent, I will use it as a chance two distinguish two options in the characterisation of the link between intentional emotional contagion and affiliative behaviour.

The first option is to think of Higgins' 'moral module' as a mechanism that underlies MME in its entirety. This would mean that Higgins is hardwired so as to automatically experience an urge to engage in affiliative behaviour upon undergoing intentional emotional contagion. His 'moral module' would thus be a physiological mechanism that reliably links a certain perception (the detection of distress behaviour), with a certain emotion (distress directed at the other's distress behaviour), and a certain behaviour (motivation to engage in affiliative behaviour). This account of how intentional emotional contagion can yield affiliative behaviour is the most straightforward one, and allows us to sidestep the worry regarding how comforting behaviour can be triggered without an understanding that the target *needs* to be comforted. Indeed, if the motivation to engage in affiliative behaviour were to be *automatically triggered* by the intentional emotional contagion, no *understanding* of the situation would be required. Such a mechanism, far from being empirically implausible, would have important evolutionary advantages. Firstly, it would ensure maternal care upon the distress behaviour of offspring—an idea that is often invoked in evolutionary explanations of empathy (see: Preston and de Waal 2002, 6). Following Stephanie Preston and Frans de Waal's analysis of the evolutionary advantages of empathy, we can also note that a mechanism of this form would help to facilitate group living, insofar as it would greatly contribute to "initiate the actions of potential allies"—in particular, their consolation in post-conflict situations—as well as "terminate the actions of … conspecific attackers" (Preston and de Waal 2002, 9).

The second option is to restrict the 'moral module' to whichever mechanism triggers intentional emotional contagion upon Higgins' detection of distress behaviour, and characterise Higgins' urge to engage in affiliative behaviour as a result of his past learning experiences. In this case, what would be hardwired is merely the link between a certain perception and a certain emotion. The ultimate deliverance of this 'moral module' would not be a motivation to engage in a specific kind of behaviour, as in the previous case, but a more general urge to eliminate what is experienced as an aversive stimulus (Jane's distress behaviour). For MME to effectively take place, it is not enough to undergo this general urge, for the simple reason that the elimination of Jane’s distress behaviour could be obtained through several other means apart from affiliative behaviour—the most radical of which would be killing Jane. If the ‘moral module’ is to be constructed in this way, Higgins must additionally have an *acquired reliable disposition* to engage in affiliative behaviour in response to distress behaviour. Higgins’ previous experiences must have taught him that this is an effective way to eliminate this aversive stimulus. Unless Higgins has this stable disposition as a result of his past learning experiences, he cannot count as a subject of MME (see also section 5.3.2). If Higgins reliably chooses to kill all individuals who trigger his intentional emotional contagion, then he would perhaps still be a moral subject, but the moral emotion that underlies his behaviour would not count as MME as I have chosen to understand it.

What might lead Higgins to prefer the affiliative method for eliminating the aversive stimulus over others, given that he lacks mindreading abilities? The response should be obvious enough if we note, as we saw in 5.1.3, that emotions and other motivations never occur in isolation, but in individuals who have many other behavioural dispositions, as well as beliefs, character traits, and so on. Thus, Higgins might have an acquired or hardwired aversion to behaving aggressively towards humans, or he might simply have a mild character. These sorts of intervening factors could have propitiated the acquisition of the stable behavioural disposition that MME requires. This second way of characterising the ‘moral module’ thus allows for differences between members of a same species, since the presence or otherwise of MME (and other moral motivations) may be modulated by factors such as the personality of each individual.

For present purposes, either of the two ways of characterising Higgins’ ‘moral module’ is equally valid, and there are no *prima facie* reasons to prefer one to the other. What is important is that MME requires intentional emotional contagion to *reliably* trigger an urge to engage in affiliative behaviour, but it doesn’t really matter whether this urge is a result of genetic determination or of past learning experiences. Additionally, while such an urge must be reliably triggered by intentional emotional contagion, it does not have to *infallibly* result in affiliative behaviour. There may be other intervening factors at specific moments in time that prevent this outcome from occurring, while not necessarily precluding the presence of MME. Consider the following example: I see you crying and, under normal circumstances, I would be inclined to comfort you. However, I have just stubbed my toe and, worried that I may have fractured it, am currently on my way to the doctor. This doesn't mean that I'm not a caring person. It's simply the case that, on this particular occasion, there is an intervening factor (my stubbed toe) that prevents what would be my normal reaction to your crying from occurring.

For MME to take place, then, we need our creature C to undergo intentional emotional contagion upon the detection of distress behaviour, and for this intentional emotional contagion to reliably trigger an urge to engage in affiliative behaviour directed at the target. While the existence of *intentional* emotional contagion has, so far, not been empirically explored, we do have evidence of a tendency, in a number of nonhuman species, to react with affiliative behaviour in response to distressed individuals. In particular, all the cases of post-conflict third-party affiliation, or any other form of *consolation*, that have been reported could have been caused by a mechanism similar to MME. As the reader may recall (see section 2.1.2.3), there is evidence of consolation in chimpanzees (de Waal and van Roosmalen 1979; Kutsukake and Castles 2004; Fraser, Stahl, and Aureli 2008), gorillas (Cordoni, Palagi, and Tarli 2006), bonobos (Palagi, Paoli, and Tarli 2004; Palagi and Norscia 2013; Clay and de Waal 2013), Tonkean macaques (Palagi et al. 2014), Asian elephants (Plotnik and de Waal 2014), domestic dogs (Cools, Van Hout, and Nelissen 2008), horses (Cozzi et al. 2010), wolves (Palagi and Cordoni 2009), prairie voles (Burkett et al. 2016), rooks (Seed, Clayton, and Emery 2007), and ravens (Fraser and Bugnyar 2010). All this, together with the evidence we saw in sections 5.2.1 and 5.2.2, strongly suggests that MME is an empirically plausible mechanism.

5.3. WHAT MAKES *MINIMAL MORAL EMPATHY* MORAL?

My argument, so far, has had two parts. In the first part, I have shown that the philosophical intuitions behind the common assumption that morality requires mindreading can all be brought into question, and in the process of doing so, I have constructed a minimal account of empathy that would not require mindreading capacities. In the second part, I have analysed the cognitive requisites of such a form of empathy and reviewed the empirical evidence that has been gathered until now and suggests that MME may exist in nature. I will now offer, as the final part in my argument, a series of positive reasons for classifying MME as a *moral* emotion. This will be a crucial point in my argumentation, for my ultimate aim is to show that morality can exist in the absence of mindreading capacities. The reasons I will put forward will all sound familiar to the reader, as they have implicitly shaped the considerations I have made until now. Still, it is very important to make them explicit at this point, in order to avoid the charge of begging the question.

Arguing for the moral character of a motivation is a very difficult endeavour, due to the fact that there are many different, and often incompatible, accounts of what morality is. It would thus be very hard, if not impossible, to give an explanation of what makes MME moral that would satisfy everyone. I will not attempt to provide an irrefutable defence of the morality of MME. Instead, my aim in this section will be a more humble one. I want to offer five different considerations that can be made about MME, which taken together build, I believe, quite a compelling case for its moral character. These are the following: (1) MME entails a sensitivity to a morally relevant property; (2) MME tracks a moral proposition; (3) MME resembles our moral phenomenology; (4) MME is a (morally right) reason, not a cause; (5) MME can be externally evaluated. All of these considerations are relatively independent from each other, and can stand or fall on their own. My aim will nevertheless be to persuade the reader that taken as a whole they are convincing enough to shift the burden of proof to whoever still wishes to deny that MME is a moral emotion.

5.3.1. MME ENTAILS A SENSITIVITY TO A MORALLY RELEVANT PROPERTY

The first reason that justifies the claim that MME is a *moral* emotion is the fact that the possession of MME entails having a sensitivity to a morally relevant property of situations. That is, even though Higgins cannot conceptualise and attribute to others the mental state that underlies displays of distress behaviour, his emotional contagion occurs in the presence of others in distress. This is important insofar as distress, as an affective state with extreme negative valence, is a *morally relevant* feature of situations. This is something about which most (if not all) theories in normative ethics will agree. From the point of view of normative ethics, whether a situation involves an individual in distress, or whether an action produces distress in others or not, is something that must generally be taken into account when deciding how to act or how to evaluate the actions of others. This is most obvious in the case of consequentialism, and especially, utilitarianism, where what determines whether a course of action is the morally right one is, precisely, the amount of pleasure and pain it produces. Within this framework, distress is considered intrinsically bad in itself, and thus morally relevant.

Distress is also morally relevant, if somewhat more subtly, in the framework of other theories. In the case of virtue ethics, for instance, an action is only virtuous when it is the expression of a certain character trait, and has been performed at the right time, in the right place, to the right person, etc. One can only have the virtue of honesty if one tells the truth consistently, but being honest also requires being sensitive to when it is the right moment to tell the truth and how the truth should be told. Whether one's interlocutor is distressed at a particular moment, or will be distressed as a result of our telling the truth, is something that will generally have to be taken into account for one's action to be an expression of the virtue of honesty. And this is not the case for every possible feature of situations. Whether our interlocutor is tall or short is not, generally speaking, something that should be taken into account. While distress is a morally relevant feature of situations within this framework, height is not.

Kantians are perhaps the only philosophers who might disagree here, for distress is not considered a moral phenomenon within the deontological framework. If one has a duty not to lie, then one should never lie, regardless of the circumstances and the possible consequences of doing so. And yet, while distress is *prima facie* not morally relevant within this framework, one could make the case that distress is *indirectly* morally relevant, or that it has a *secondary* moral relevance. For instance, according to deontologists, one has the duty not to kill, which means one should never kill. But suppose Jim were in a situation where he had no other choice but to kill Peter. Suppose, then, that the only choice that Jim could make were between killing Peter in a slow, painful manner (thus causing him severe distress), and doing so in a quick and painless way. If we were to ask a Kantian what the morally right thing for Jim to do is, the Kantian would have to choose between saying that both options are equally bad and it doesn't matter which one Jim chooses—which would be very counter-intuitive, to say the least—, or saying that Jim should choose to kill Peter painlessly, which would ultimately mean that pain (and, by extension, distress) is a bad-making feature of situations; that it has moral relevance, even if only indirectly.

It seems fairly uncontroversial, then, to say that distress is a bad-making feature of situations, since those situations that involve distress will, *ceteris paribus*, be considered morally worse than those that don't. This means that distress is morally relevant. The type of sensitivity that Higgins has to this morally relevant feature of situations fits in with the one Rowlands' (2012b) requires of moral subjects, for he says that "[m]oral subjects are ones who are sensitive to the good- and bad-making features of situations in the sense that they entertain intentional content emotionally" (Rowlands 2012b, 230). Accordingly, Higgins possesses an *emotional sensitivity* to this morally relevant feature, since he not only feels distressed whenever he sees others in distress, but also *experiences* their distress behaviour *emotionally* as distressing, for the distress that results from his emotional contagion is *intentionally directed* at their distress behaviour. Higgins may also experience the vacuum cleaner emotionally as distressing, but this sensitivity would not be a moral one because vacuum cleaners are not morally relevant features of situations.

Higgins has an emotional sensitivity to distress, and this sensitivity is not accidental or contingent, but rather, grounded in what Rowlands would call a "moral module," that is, a psychological mechanism, such as a perception-action mechanism (Preston and de Waal 2002), that entails that Higgins' emotional sensitivity is *reliable*. The reliability of the operations of this mechanism grounds the truth of counterfactual statements that we can use to describe Higgins' sensitivity, such as: *ceteris paribus*, if the situation did not involve distress, then Higgins would not undergo intentional emotional contagion, and would not be compelled to engage in other-directed affiliative behaviour (see: Rowlands 2012b, 225, ff.).

A certain objection can be anticipated at this point. If a behaviour reader can detect and reason about behavioural cues, but not understand behaviour in terms of the mental states that underlie it, how can we say of Higgins that he possesses a sensitivity to (at least) one bad-making feature of situations? Is it not precisely the *phenomenology* underlying displays of distress behaviour, that is, *what it feels like* for the subject in distress, that which makes it an objective bad-making feature of situations? This is right, but it need not be a problem. Although Higgins lacks mindreading capacities, he is sensitive, not just to distress behaviour, but to distress itself. Indeed, distress behaviour is a *reliable marker* of distress, and it is sufficient for Higgins to be sensitive *to* distress in others (in the sense that, when faced with distress in others, he undergoes intentional emotional contagion) but we do not need him to be sensitive *that* it is, in fact, a case of distress as an unpleasant mental state. That is, the fact that he reliably experiences (intentional) distress when in the presence of another creature in distress is enough to say that he is sensitive *to* the bad-making feature of situations that is the suffering of others. The fact that he doesn't *represent* distress in others as an unpleasant mental state doesn't preclude him from having such sensitivity.

The importance of the conditions we put forward in section 5.2.3, when discussing whether the subject of MME should possess the concept DISTRESS BEHAVIOUR, should now become clear. As the reader may recall, we determined that Higgins should be able to (1) identify and reidentify instances of distress behaviour, (2) detect distress behaviour in different individuals, (3) represent different properties of individuals (so that he is not merely a detector of distress-behaviour-situations, but understands distress behaviour as a property of individuals), and (4) detect distress behaviour in its different forms and classify them using the same stable representation—the behavioural abstraction 'distress-behaviour.' All of these conditions are important because they allow us to say that while Higgins may not be able to understand *that* others are distressed, he is sensitive *to* distress in others. Possession of the behavioural abstraction 'distress-behaviour,' regardless of whether it counts as a concept or not, grants Higgins this sensitivity because it allows him, insofar as distress behaviour is a reliable marker of distress, to consistently react to the presence of distress in others.

Another objection emerges at this point: if this is enough to say that Higgins has a minimal form of morality, why not establish that he possesses a *minimal form of mindreading*? There would be a certain truth behind this objection, insofar as MME would not be a moral emotion were it not for the fact that it reliably occurs when in the presence of an individual in a certain *mental state* (distress). MME is what Stephen Butterfill and Ian Apperly call a "theory of mind ability," for it is "an ability that exists in part because exercising it brings benefits obtaining which depends on exploiting or influencing facts about others' mental states" (Butterfill and Apperly 2013, 607). Possessing MME, that is, having the ability to detect distress behaviour in others and respond to it with care, can generate benefits to an animal community, insofar as it can ensure maternal care and facilitate group living. These benefits obtain *in virtue of the fact* that MME is triggered when in the presence of, and generates an urge to act upon, others' mental states. In this sense, and only in this sense, it is a "theory of mind ability." At the same time, however, MME doesn't require, to follow Butterfill and Apperly's terminology, any "theory of mind cognition," which would involve "representing mental states or processes as such" (Butterfill and Apperly 2013, 607). This qualification is not a trivial one. It means that there are certain things that Higgins can't possibly do. He can't tell the difference between real distress and feigned distress, and acknowledge that only the former is a bad-making feature of situations. He can't understand how his behaviour can influence the mental states of others, or evaluate different courses of action in light of the effect they may have on others' mental lives. He can't justify his comforting behaviour or explain why it was right of him to respond that way to the other's distress. All of this means he can't be held responsible, praised or blamed, for his comforting behaviour towards others because he lacks the requisite form of understanding. These differences between Higgins—a 'mere' moral subject—and someone with (a certain degree of) moral agency mark an important boundary. Whether or not Higgins' "theory of mind ability" should be labelled 'minimal mindreading' is largely a matter of stipulation. The important point, for my purposes, is that no "theory of mind cognition" *whatsoever* is required for Higgins' to behave on the basis of MME, and thus to qualify as a moral subject.

5.3.2. MME TRACKS A MORAL PROPOSITION

As we saw in section 5.2.3, for an affective state to be an emotion, and not simply a mood, it must contain factual content. This means that there must be a state of affairs that the emotion is directed at. In the case of MME, the factual content can be specified by use of the proposition "This creature is displaying distress behaviour." Emotions are also characterised because they contain evaluative content. This means that the state of affairs that the emotion is directed at is evaluated in a certain way. For instance, if an individual is sad about something, this means she considers it as something she *should* be sad about; something that *warrants* sadness. In the case of *moral* emotions, this evaluation takes a moral character—the object or state of affairs that the emotion is directed at is evaluated as morally good/bad/right/wrong/etc. In the case of MME, the evaluative content required can be specified by the proposition "This creature's distress is bad."

The requirement that moral emotions contain morally evaluative content has often been interpreted as the idea that a being cannot possess a moral emotion unless she is capable of engaging in explicit moral judgements. This idea is present, for instance, in the work of Beth Dixon, who states that empathy won't be a moral emotion unless it "involves the evaluative judgement that the object's distress is bad" (Dixon 2008, 133). She claims that animals can't possibly entertain the evaluative judgements that moral emotions require, because "[t]his cognitive-evaluative intentional content characteristic of moral emotions is not susceptible to … quantitative gradualism" (Dixon 2008, 141). This means that "morally laden emotions are not the kinds of psychological states that can be carved up quantitatively by increments or degrees and attributed in some lesser amount to any kind of subject" (Dixon 2008, 143). Entertaining the sorts of evaluative judgements required for moral emotions is, in her view, an all-or-nothing matter. Since Higgins, being a dog, clearly lacks the requisite concepts for judging that a creature's distress is bad, this would mean, according to Dixon, that he cannot possess a moral form of empathy.

Rowlands (2012b) provides us with a way around this problem; a way of "carving up," so to speak, the required moral judgement and attributing it in a "lesser amount" to Higgins. In Rowlands framework, moral emotions must contain morally evaluative content, but it is not necessary that the individual in question be capable of *explicitly entertaining* this content. All that is required for MME to count as a moral emotion is for it to *track* the proposition "This creature's distress is bad." To understand what this means, we should begin by pointing out that moral propositions are meant to be a record of how the world is. That is, the proposition "This creature's distress is bad," if true, would mean that this creature's distress is indeed a bad thing. These propositions thus have a *descriptive* component. But moral propositions are not *merely* a record of how the world is—they also have a *motivational* component. This means that if I believe that "X's distress is bad" is true, then *ceteris paribus*, upon seeing that X is in distress, I will be moved to help or comfort X. With this in mind, we can establish that Higgins' distress at others' distress (behaviour) tracks the proposition "This creature's distress is bad" because of the phenomenal character that Higgins' emotion takes. On the one hand, Higgins reliably feels distressed whenever he sees others displaying distress behaviour, where his distress is not merely triggered by, but also directed at, their distress behaviour. This means he experiences distress in others as something negative, unpleasant. The *descriptive* aspect of the moral proposition "This creature's distress is bad" is thus present in Higgins' aversion to others' distress (behaviour). On the other hand, Higgins' emotion also has the adequate *motivational* component to it, insofar as, built into his distress, is an urge to engage in affiliative behaviour directed at the target. Even though Higgins cannot entertain the moral proposition that MME tracks, it is implicit in the phenomenal character of Higgins' emotion, in the sense that Higgins reliably experiences others' distress (behaviour) as something negative, and as something he wishes to eliminate.

The phenomenal character that Higgins' emotion takes thus warrants us describing it as tracking the proposition "This creature's distress is bad." The fact that MME tracks the proposition "This creature's distress is bad," in turn, means that there is a certain truth-preserving relation between Higgins' emotion and this proposition (see Rowlands, 2012b, Chapters 2, 9). This truth-preserving relation can be best explained by looking at the errors that can be involved in an emotion like MME. We have already mentioned them, but it's worth going back to them at this point. As we've seen, there are two ways in which MME can *misfire*. Suppose, first, that Higgins sees Jane vocalising and gesticulating in a way that resembles distress behaviour, and his moral module triggers MME. But suppose that Jane is not actually distressed. She is rehearsing her part in a play. In this case, Higgins' emotion is based on a factual proposition—"This creature is displaying distress behaviour"—that is false (insofar as feigned distress behaviour is not actual distress behaviour). In Rowlands' terminology, we would say that Higgins' emotion is *misplaced*.

Now, imagine that Jane actually *were* distressed, and Higgins, motivated by his MME, walks up to her and starts nuzzling her. Let us now suppose that even though Jane is, in fact, distressed, her distress is not the result of any wrongful, or otherwise negative, event in her life. She is merely reading a sad novel, and her distress has been caused, and is directed at, this work of fiction, and not any state of affairs in her ‘real’ life. Or, alternatively, Jane has just learnt that she is getting sent to prison on account of a cold-blooded murder she committed, and that is why she is crying. In cases like these, Jane’s distress would, at least arguably (and depending on the normative ethics we subscribe to), not be a *bad* thing. The moral proposition “Jane’s distress is bad” is, in these types of circumstances, false. We would then say, following Rowlands, that Higgins’ emotion is *misguided*, because it is based on a false evaluative proposition. There is thus a truth-preserving relation between Higgins’ emotion and the moral proposition “Jane’s distress is bad” such that, if Higgins’ emotion is not misguided, then this proposition *must* be true. The truth of this proposition is *guaranteed* by the non-misguided status of Higgins’ emotion. This truth-preserving relation means that MME counts as a morally laden emotion within Rowlands’ (2012b) framework:

> An emotion, E, is *morally laden* if and only if (1) it is an emotion in the intentional, content-involving, sense, (2) there exists a proposition, *p*, which expresses a moral claim, and (3) if E is not misguided, then *p* is true. (Rowlands 2012b, 69, his emphasis)

The idea of MME tracking a moral proposition further allows us to say that Higgins’ emotion would be moral even if it were misguided or misplaced. Higgins doesn’t have to be infallible at detecting distress behaviour. And it doesn’t necessarily have to be the case that Jane’s distress is indeed a bad thing. This accommodates the intuition that caring or comforting behaviour is moral even if it has been caused by an emotion that has misfired (see section 5.1.4). What is crucial is not for Higgins to be an infallible moral creature, but for his behaviour to be motivated (at times) by an emotion that tracks a moral proposition. And we know that MME tracks the moral proposition “This creature’s distress is bad” because it has the adequate phenomenal character and because, were it not to be misguided, the truth of this proposition would be guaranteed.

5.3.3. MME RESEMBLES OUR MORAL PHENOMENOLOGY

MME resembles the phenomenology of what us humans often experience when being moved by another's distress to help or comfort her. While there are many different kinds of reasons that can move us to attempt to alleviate an individual's distress, it is frequently the case that, when confronted with an individual in distress, we experience it as something intensely unpleasant, and built into this unpleasantness is an urge to comfort or help that individual. This is perhaps more common in those cases where our reaction is very primary, almost instinctive. A good way to see this is to think about how quickly we are moved to comfort a child who has just tripped, fallen over, and started crying. In these sorts of cases, explicitly attributing a mental state to the target, or thinking "This person's distress is bad" is not usually a step along the way—we may even be considered callous if we need to stop and make these sorts of considerations before helping someone in distress.

We can, of course, engage in moral reflections *post hoc* ("It was good of me to help her, after all, her distress was a bad thing"), and this is relevant when it comes to granting us moral agency. However, since we are only considering granting moral subjecthood to Higgins, and not moral agency, we are solely focusing on the motivation behind his behaviour, for a 'mere' moral subject is an individual who behaves on the basis of moral motivations. Granting moral subjecthood to Higgins strictly depends on the kinds of motivations that underlie his behaviour. Given this fact, acknowledging that his motivation to comfort Jane is one that we can relate to, one that is phenomenologically similar to what we often experience in cases where we commonly consider our behaviour to be moral, is an important *prima facie* reason for considering that the motivation behind Higgins' behaviour is moral too.

5.3.4. MME IS A (MORALLY RIGHT) REASON

In section 4.3, we saw an experiment in which ants reliably freed non-anesthetised conspecifics who were trapped in a snare (Nowbahari et al. 2009). According to the authors themselves, the ants' rescue behaviour was probably triggered by an "eliciting stimulus"—presumably a pheromone—that was "actively produced" by the entrapped conspecific (Nowbahari et al. 2009, 3). While merely appealing to such a "chemical call for help" probably doesn't suffice to explain the precisely targeted helping behaviour observed in the ants (Nowbahari et al. 2009, 3), let's assume, for explanatory purposes, that it was the only mechanism underlying the ants' performance. Let's suppose, then, that the 'chemical call for help' automatically triggered the ants' rescue behaviour, and did so without the mediation of any conscious mental state. Following Rowlands (2012b), we would say that the 'chemical call for help' was the *cause* of the ants' behaviour. In contrast, MME can be thought of as not only the *cause* of Higgins' behaviour, but also the *reason* behind it, since it doesn't merely trigger the behaviour, but rather, *motivates* it by way of its intentional content (Rowlands 2012b, 35).

The notion of *reason* is used in three different contexts, depending on whether we are concerned with deliberation, explanation or justification (Lenman 2011). In contexts of deliberation, one looks for and evaluates different reasons for performing an action. For instance, if Paula is considering quitting smoking, she may engage in a deliberation where she weighs different considerations that speak in favour of or against smoking. Thus, she may consider the amount of money she spends on cigarettes and the effects smoking has on her health as reasons to quit smoking, while simultaneously considering the amount of pleasure she gets from this activity as a reason not to quit. When stating that MME is a reason for Higgins' behaviour, I am not thinking about the context of deliberation. Since Higgins—we are supposing—lacks metacognition, mindreading capacities, and moral judgement, he cannot engage in any such deliberation where he might evaluate different reasons in favour of and against comforting Jane. While we cannot speak of deliberation in reference to Higgins, we can nevertheless speak of MME as a reason that *explains* and one that *justifies* his behaviour.

When we attempt to *explain* someone's behaviour, we can invoke the notion of reasons, not to refer to the considerations that could have spoken in favour of that individual behaving that way, but to refer to what *actually motivated* the agent to behave in that way. In this context, reasons are usually thought of as belief-desire couplings. For instance, we may find that Paula's decision to quit smoking is explained by her *belief* that smoking is bad for one's health and her *desire* to stay healthy. This belief-desire coupling is the *reason* that motivates her quitting. MME, as we saw, is an emotion that involves both factual and evaluative content. This allows us to say that it provides a reason for Higgins' comforting behaviour that "mirrors the factual-affective motivating profile of belief-desire couplings" (Rowlands 2012b, 193). Higgins' comforting behaviour towards Jane is explained by the fact that Higgins has an aversion to distress behaviour and believes, as he sees Jane crying, that she is displaying distress behaviour. The factual content that is involved in his MME ("This creature is displaying distress behaviour"), coupled with the evaluative content that it tracks ("This creature's distress is bad"), motivate his affiliative behaviour. Because MME motivates Higgins' behaviour in virtue of its content, it counts as a reason for his behaviour, and not a mere cause.

MME is a reason, not just because it explains Higgins' behaviour, but also because it *justifies* it. When speaking of the justification behind Higgins' behaviour, I'm not, of course, thinking of any justification that Higgins can verbalise, nor even one that he can reflect upon in any way. This is not just due to the fact that Higgins doesn't possess linguistic abilities, but also because he lacks mindreading skills, moral judgement, metacognition, etc. However, the fact that Higgins himself can't justify his behaviour doesn't mean that it isn't justifiable. We, as full-blown moral agents, equipped as we are with all the cognitive mechanisms that Higgins lacks, can analyse the reason behind Higgins' behaviour—the *motivating* or *explanatory* reason—and ask whether it was a *sensible* reason for comforting Jane, whether it actually spoke *in favour of* this behaviour. Being distressed *about* Jane's distress behaviour is a *justifying* reason for engaging in affiliative behaviour, insofar as this sort of behaviour will reliably contribute to ending the stimulus that Higgins finds aversive in the first place. If Higgins were positively sensitised to distress behaviour, instead of negatively sensitised, we would say that he has no justifying reasons to engage in affiliative behaviour, for by comforting Jane he would be eliminating a stimulus that he is, in fact, enjoying. MME acquires the character of a *justifying* or *normative* reason, because it speaks *in favour of* comforting Jane, and renders this behaviour sensible (see, e.g.: Dancy 2000, 1, ff.).

Not only does MME justify Higgins' comforting behaviour from a *practical* perspective, it also does so from a *moral* one. Distress, as we've seen, can plausibly be regarded as a bad-making feature of situations. The fact that Jane is distressed and that her distress is indeed a bad thing turns MME into what Higgins *should* feel, given the circumstances. Higgins not only comforts Jane; he does so, we are supposing, on the basis of a motivation that has an adequate phenomenal character. Even though Higgins can never realise this, we, as moral agents with moral judgement capacities, can assert that MME was the *morally right* thing to feel in those circumstances, for when he undergoes MME, Higgins is experiencing as bad something that is, in fact, bad. This allows us, as external evaluators, to say that he comforts Jane *for the right reasons*, which, in turn, can serve as grounds to evaluate Higgins' MME-based behaviour as morally right. The fact that Higgins' behaviour can be labelled 'morally right' because it has been motivated by reasons of the appropriate sort means that Higgins is a moral subject, even though his different cognitive deficiencies prevent his behaviour from being praiseworthy.

5.3.5. MME CAN BE EXTERNALLY EVALUATED

When it comes to evaluating an individual's behaviour or character, we can, following Julia Driver (2006), distinguish two different approaches: 'evaluational internalism' and 'evaluational externalism.’ According to ‘evaluational internalism,’ “the moral quality of a person’s action or character is determined by factors internal to agency, such as a person’s motives or intentions.” In contrast, ‘evaluational externalism’ is “the view that the moral quality of a person’s action or character is determined by factors external to agency” (Driver 2006, 68). In the previous section, I sketched an internalist account of the morality of Higgins’ behaviour, according to which, Higgins’ comforting behaviour qualified as morally right because it had been motivated by a morally right reason. MME can also be subjected to an externalist account of its moral quality, such as the one offered by Driver herself (which Rowlands also follows—see: Rowlands 2012b, chap. 9).

Driver understands character traits as psychological dispositions to feel, act or behave in a certain manner (Driver 2006, 68). Possessing MME means that Higgins has a reliable disposition to feel distress and engage in affiliative behaviour in response to the distress (behaviour) of others. Under Driver’s account, MME would thus count as a character trait of Higgins. Moreover, we have some evidence that affiliative behaviour, at least in the case of mammals, effectively contributes to alleviating others’ distress (Kikusui, Winslow, and Mori 2006; Fraser, Stahl, and Aureli 2008; Clay and de Waal 2013; Smith and Wang 2014; Burkett et al. 2016). Since distress is a bad-making feature of situations, MME will systematically produce good consequences, for it will tend to diminish the distress of the target, thus resulting in a decrease in the amount of bad in the world. Within Driver’s framework, this means that MME counts as a virtue, which she defines as “character traits that systematically produce more actual good than not” (Driver 2006, 68). Under her de-intellectualised theory of virtues, Higgins is a virtuous individual.

Even if we don’t subscribe to Driver’s account of virtues, the fact that MME systematically produces good consequences allows us to evaluate Higgins’ MME-based behaviour, from an objective consequentialist perspective, as morally right. While we cannot praise Higgins for his behaviour, we can say that, whenever he comforts others in distress, “the world is—temporarily, perhaps even momentarily—a better place,” and so, that it is “a good thing that the world contains a subject like this, an individual who acts in this way" (Rowlands 2012b, 254). That this sort of external evaluation is possible, counts in favour of considering that Higgins is a moral subject. As the reader may recall, Rowlands (2012b) characterised a *minimal moral subject* in the following way:

> X is a moral subject if X possesses (1) a sensitivity to the good- or bad-making features of situations, where (2) this sensitivity can be normatively assessed, and (3) is grounded in the operations of a reliable mechanism (a "moral module"). (Rowlands 2012b, 230)

In section 5.3.1, we already saw that possessing MME entails having a sensitivity to (at least) one bad-making feature of situations. I have also discussed (in sections 5.2.4 and 5.3.1) how such sensitivity can be regarded as being grounded in the operations of a 'moral module.' The possibility I have sketched in this section and the previous one, of subjecting Higgins' MME to a normative evaluation—either internal or external—, means that his sensitivity can be normatively assessed. Therefore, he fulfils the sufficient conditions put forward by Rowlands to be considered a moral subject.

5.4. SUMMARY

In this chapter, I have defended the main thesis of this dissertation, which is the idea that morality can obtain in the absence of mindreading capacities. I began by looking at the different arguments that philosophers have used to defend the idea that moral behaviour requires mindreading capacities. I have argued that, while Cheryl Abbate (2014) may be right in requiring mindreading for moral responsibility, this doesn't preclude the existence of moral subjects that lack this capacity. I have shown that, contrary to what Florian Cova (2013) argues, moral responsibility cannot be separated from moral judgement, and that the latter is not necessary to give an account of minimal forms of care based on empathy. I have criticised Nichols' (2001) arguments in favour of a necessary connection between mindreading and altruistic motivation, and given the reasons why de Boer's (2011) account of the "basics of morality" is intellectualistic to the extent that renders it implausible. Along the way, I introduced the notion of MME, which was then systematically defined, and the reasons for thinking that it is an empirical possibility were offered. In the final section, I gave five different accounts of why MME should be understood as a *moral* emotion. I have argued that it entails a sensitivity to a morally relevant feature of situations, that it tracks a moral proposition, that it has a similar phenomenology as some of our moral motivations, and that it allows for two different normative evaluations—internal and external. If my arguments have been correct, there is at least one moral motivation that is within the reach of non-mindreading animals.

6. CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

Could animals behave morally if they can't mindread? I have tried to show that there are important grounds for considering that morality, understood as the capacity to behave on the basis of moral motivations, can not only obtain in the absence of moral judgement and moral responsibility, but also in individuals who lack mindreading abilities. We've seen that the arguments that have been adduced to support a necessary connection between moral behaviour and mindreading capacities do not stand up to scrutiny. I have further shown that a minimal form of empathy that would not require mindreading capacities and would nevertheless count as a moral emotion is a conceptual possibility. While MME is intended as a theoretical construct that shows that morality and mindreading can be brought apart, there are some reasons to consider it an empirical possibility. In particular, we've seen that there is already a significant amount of empirical evidence that strongly suggests that the components that make up MME are present in several nonhuman species. What's more, this evidence has been, for the most part, gathered in the last decade, and the focus of these studies has been on a small range of species (mostly mammals), so it's likely that more evidence will appear in upcoming years.

The notion of MME can be used to answer the question we began with, namely, whether behaving morally necessarily requires the possession of mindreading capacities. MME points us in the direction of a negative answer to this question, as it shows that moral behaviour is conceivable in the absence of mindreading capacities. But this is not the only theoretical implication that derives from this notion. In particular, this concept can contribute to advancing the three debates we saw in chapters 2, 3, and 4: the animal morality debate, the animal mindreading debate, and the debate on the nature of empathy.

In the animal morality debate, a discrepancy exists between the growing amount of empirical evidence that suggests that (some) animals may be capable of behaving morally, and the blanket denial of animal morality that is defended, on *a priori* grounds, by many philosophers. This discrepancy stems, at least partly, from the pervasiveness of the amalgamation fallacy. The best antidote to this situation, I have argued, is to adopt Rowlands' (2012b) conceptual framework, incorporating the category of moral subjecthood into debates on animal morality. The notion of MME helps to reinforce this framework in two distinct ways. One the one hand, MME is the result of the application of Rowlands' theory to the analysis of a particular moral emotion. As such, it shows that this framework not only works in abstract terms, but can also help make sense of much empirical evidence, and shine a light on debates on empathy as a moral motivation, which tend to lack conceptual clarity. On the other hand, MME helps to develop this framework, by showing that moral subjecthood is indeed independent of higher-order thinking, not just because it doesn't require metacognition (as argued by Rowlands 2012b), but also because it can obtain in the absence of mindreading capacities—the other great class of higher-order thinking. MME thus contributes to the animal morality debate by exemplifying how moral motivations can be stripped from over-intellectualistic requirements while retaining the core characteristics that make them moral. By shifting the focus away from the most intellectually demanding forms of morality, MME helps move our attention from the question 'Are animals moral in the same sense that humans at their most lucid and self-reflective can be?' to the question 'Can some of the motivations behind some animals' behaviour be reasonably considered moral?' This second question has a greater philosophical and scientific interest than the first one, insofar as it doesn't have such a clear and obvious answer.

The notion of MME as I have developed it also adds to the mindreading debate. Obviously, it does not bring us any closer to a solution to the question of whether any nonhuman animal can engage in mindreading, but it contributes to this debate indirectly, insofar as it impels us to stop and reflect on the importance of answering this question. Mindreading is often considered a crucial cognitive ability, not in itself, but because of the advantages it grants when it comes to predicting, explaining, and understanding the behaviour of those in our immediate environment. Additionally, it is thought crucial for morality, for many different reasons, as we've seen. My discussion of the possibility of moral behaviour existing in the absence of mindreading capacities adds to a number of dissenting voices who have recently begun to question the significance of this ability (see, e.g.: Andrews 2012b; Leudar and Costall 2011). MME is an example of how social behaviour that isn't triggered by mindreading capacities can nevertheless possess qualities that have been traditionally thought of as exclusively human. In this sense, it contributes to undermining the centrality of mindreading capacities. This is also important because the animal mindreading debate is stuck at an impasse with no signs of resolution in the near future. As we saw, what little evidence we have applies to a very small number of species and is viewed as deeply problematic by several prominent researchers. MME shows that the case for animal morality should be considered independent of the case for animal mindreading, which means that those interested in the former debate should not worry about the discouraging results in the latter.

With respect to the empathy debates, there is an immense, and growing, amount of literature dedicated exclusively to how we should characterise the very notion of empathy. I have not intended to add one further definition of empathy to the innumerable quantity that already exists. In fact, I sense that there is no 'one' definition of empathy to be found. Rather, this should be properly viewed, in my opinion, as an umbrella term for a number of different natural phenomena that share a certain 'family resemblance.' I am not so much interested in what empathy actually is, but in what there is to learn from the countless debates that revolve around this topic. Within discussions on the nature of empathy as a moral motivation, there is one particular claim that all authors seem to agree on, despite characterising empathy in many varied ways. This is the idea that, for empathy to count as a moral motivation, mindreading capacities are required. To determine whether something that looks like empathy-motivated moral behaviour actually qualifies as such, it is usually thought, we have to look at the underlying cognitive architecture and see whether mindreading capacities are present. If they are not, so the story goes, then it's probably a case of aversive arousal, a purely selfish reaction that doesn't qualify as moral. MME is an important contribution to this debate because it brings into question the widely accepted dichotomy between moral motivation and aversive arousal. I have shown that aversive arousal, contrary to common belief, can count as a moral motivation so long as certain conditions are met, amongst which, crucially, mindreading capacities are not included. MME thus lowers the standard for empathy to count as a moral motivation, which not only opens the door to the possibility of nonhuman moral behaviour, but also to the idea that humans with cognitive deficiencies that may come from a lack of development, a mental illness, or any other impairment, may nevertheless be moral subjects.

While the notion of MME allows us to advance these three debates, there are also some important questions that are immediately opened up by it. Perhaps the most obvious of these is the question of whether it is possible to obtain conclusive empirical evidence of MME, and if so, what a decisive experiment might look like. Further, given the fact that not every animal may be capable of emotional contagion, and that MME is but one example of a minimal moral emotion, we can also ask whether any other moral emotions can exist in the absence of mindreading capacities. Given that morality may obtain in the absence of mindreading capacities, we can also consider whether minimal moral emotions could have played any role in the evolutionary emergence of mindreading. And lastly, for those of us interested in animal ethics, an additional question arises regarding the ethical consequences that might follow from adopting the notion of minimal morality, and in particular, whether there are any special duties that moral agents would hold towards minimal moral beings. In the following sections, as a final step in this dissertation, I will have a look at each of these four questions, and introduce some ideas as to how we could go about answering each of them. This will give us some clues to determine where the research could go from here.

6.1 Empirical evidence for MME?

A subset of the empirical evidence we saw in chapter 2 can be explained by postulating the presence of MME. In particular, all instances of third-party post-conflict affiliation that have been observed in the wild, as well as all other cases of apparent consolation behaviour, may have been triggered by a mechanism similar to MME, for they all involved affiliative behavior by an uninvolved bystander towards a distressed individual. As the reader may recall, this behaviour has been observed in chimpanzees (de Waal and van Roosmalen 1979; Kutsukake and Castles 2004), bonobos (Clay and de Waal 2013; Palagi and Norscia 2013), gorillas (Cordoni, Palagi, and Tarli 2006), dogs (Cools, Van Hout, and Nelissen 2008), wolves (Palagi and Cordoni 2009), horses (Cozzi et al. 2010), Tonkean macaques (Elisabetta Palagi et al. 2014), Asian elephants (Plotnik and de Waal 2014), rooks (Seed, Clayton, and Emery 2007), and ravens (Fraser and Bugnyar 2010). In addition to these observational studies, the performance of animals in some experiments may also be accounted for by the presence of MME. The experiment by Deborah Custance and Jennifer Mayer (2012), where dogs engaged in affiliative behaviour towards humans who were feigning distress behaviour, as well as the recent experiment by James Burkett *et al.* (2016), where prairie voles increased their allogrooming behaviour when they were paired with conspecifics in distress, could both be the result of a mechanism similar to MME operating in these two species.

In several of the experiments in the altruism cluster (section 2.1.1.3), the experimental subjects were faced with a conspecific in distress, but could not engage in affiliative behaviour due to the existence of a physical obstacle of some sort. Nevertheless, if they were given the chance, the animals often chose to extinguish the cause of the other's distress (behaviour), thus indirectly contributing to an alleviation of the latter. The pioneer experiment by Russell Church (1959), where rats refrained from pressing a lever to obtain food when they saw that it caused another rat to receive an electric shock, was the first of these experiments. Similar results using this paradigm were obtained with pigeons (Watanabe and Ono 1986), and rhesus monkeys (Masserman, Wechkin, and Terris 1964; Wechkin, Masserman, and Terris 1964). Further experiments with rats have shown that they will reliably press a lever to lower a conspecific suspended from a harness (Rice and Gainer 1962), that they will choose to obtain food from a lever that terminates a shock being delivered to a conspecific (Greene 1969), that they will refrain from walking down the arm of a T-maze that results in a shock delivered to a conspecific (Evans and Braud 1969), that they will reliably free conspecifics from a restrainer (Bartal et al. 2011), and that they will choose to help a conspecific exit a container full of water (Sato et al. 2015). The evidence we have of emotional contagion in rats (Atsak et al. 2011; Nakashima et al. 2015) suggests that a mechanism akin to MME may be at the basis of their performance in these experiments.

While the behaviour of the animals in all these studies is *compatible* with an explanation in terms of MME, none of them provides conclusive evidence of the existence of this moral motivation. In fact, they are all compatible with an explanation in terms of 'mere' aversive arousal, that is, one in which the subjects decide to eliminate the target's distress (behaviour) merely because they find it unpleasant, where the content of the subject's distress is the feeling of unpleasantness itself, or where there is no clear self-other distinction and the subject's helping or comforting behaviour is, in fact, entirely self-centred. MME, as we saw in chapter 5, is a special form of aversive arousal, with two specific components that distinguish it from an entirely selfish reaction: (1) the intentional content of the subject's distress is the other's distress behaviour, and (2) built into this distress is a specific urge to engage in affiliative behaviour. To obtain conclusive evidence of MME, we would need to distinguish 'mere' aversive arousal from the one that makes up MME. If we were to focus on component (1), this task becomes practically insurmountable, for how could we possibly access the intentional content of an animal's emotion? One possibility might be to compare her physiological and behavioural reactions when confronted with another individual in distress, and when faced with any other aversive stimulus, in hopes of finding some crucial difference that may give us a clue as to the intentional content of her distress in either case. However, it is very difficult to devise what such a crucial difference might look like.

A better alternative, I suspect, would be to focus on component (2). In instances of 'mere' aversive arousal, or personal distress, there would be no difference for the subject between alleviating the other's distress and escaping the situation. In fact, the latter might be preferred in those cases where it is less costly. On the other hand, were MME to be under operation, acting upon the other's distress (behaviour) should be the preferred alternative, because that is the object in the world that the subject's distress is directed towards. In none of the experimental studies surveyed were the animals given a choice between escaping the situation and alleviating the other's distress. Their choice was always between extinguishing the other's distress (behaviour) and continuing to hear/see/smell it. If they were to be given a choice between escaping the aversive stimulus and acting upon it to eliminate it, and if they still chose the latter, then this would give us quite strong evidence that the other's distress behaviour was, in fact, what they intended to eliminate, and not merely their own personal distress. While this would not be rock-solid, irrefutable evidence of MME—arguably, this is an ideal we should aspire to but ultimately an unattainable one—, if we were to combine it with measures to test the presence of emotional contagion, I suspect that this is as close as we could get to a proof of the existence of this moral motivation.

I am, however, wary to suggest that such a study should be performed, as I have strong reservations regarding the ethical character of this hypothetical experiment. In an ideal world, we would be able to test for a positive form of MME, that is, one in which the emotional contagion is triggered by the display of a positive emotion in another, and this generates an urge to somehow contribute to the permanence or enhancement of the target's joy. Unfortunately, it is very difficult to come up with an experiment that could test for this positive form of MME.

This is especially due to the fact that when we are faced with something that makes us happy, we may be perfectly inclined to simply leave things as they are, in a way that we're not when we're faced with something that makes us suffer. Therefore, I suspect that it would be very complicated to place an animal in a situation where she undergoes positive emotional contagion and is still moved to perform a behaviour that shows us that she wants to contribute to the permanence or enhancement of the target's joy (behaviour).

The experimental road to conclusive evidence of MME thus faces important problems. The solution may, alternatively, be to focus on observational studies. After all, in the wild animals *do* have the chance to escape the aversive stimulus that is others' distress behaviour, so the fact that animals of many different species still often choose to engage in consolation behaviour may be an indication that they are behaving on the basis of MME. This is by no means an unquestionable form of evidence, for the animals could be moved by other motivations, such as an expectation that the target may reciprocate in the future. While controlling for alternative explanations in the wild is immensely difficult, this problem can be overcome, to a certain extent at least, by tracking each individual's personal history, personality, and relationships with her peers. Careful and long-term observational studies performed in the wild or in captivity, coupled with the experimental evidence we already have, would undoubtedly be the most ethical approach to the study of animal morality. In my opinion, and this is something that I can only state here but not defend, this ethical character of observational studies trumps any advantages that experimental approaches may have. Unless we are capable of devising a way to test for a positive form of MME, the best alternative, in my view, is to pursue the observational approach.

6.2. OTHER MINIMAL MORAL EMOTIONS?

MME is an example of the application of Rowlands' theoretical framework to the analysis of a specific type of moral emotion. The existence in nature of MME, as I have defined it, is dependent upon the presence, in different species, of the components that make it up. As we've seen, there is already a considerable amount of evidence that suggests that this may be the case. And yet, it is unlikely that all candidates for moral subjecthood in the animal kingdom will possess MME—it may even turn out that MME is nothing but a theoretical construct. In order to strengthen the case for animal morality, the most natural step would now be to further develop Rowlands' framework, and show how it can allow us to construct a broader account of minimal morality that encompasses other moral motivations. I believe that this endeavour is feasible, and that there are indeed further moral emotions that could motivate the behaviour of animals who are mere behaviour readers. While for now I can only offer a brief sketch of how this project might go, I hope to show that there are no conceptual obstacles to thinking that morality in a broader sense, and not just empathy, can take place in the absence of mindreading capacities.

If we look at the phenomenal character of MME, we would classify it as a *negative* moral emotion, insofar as the core affection that underlies it—distress—has a negative valence. Experiencing MME is, on the whole, unpleasant. As we briefly saw in the previous section, a positive counterpart of MME is perfectly conceivable. We could imagine that Higgins, apart from having the capacity to detect distress behaviour, were capable of detecting joy behaviour and categorising as such. We could imagine his moral module ensuring that he undergoes emotional contagion, not only when witnessing distress behaviour, but also when detecting joy behaviour. Higgins would not only feel sad when he sees others displaying distress behaviour, but would also feel happy when he sees others displaying joy behaviour, where this happiness were not merely triggered by, but also directed at, said stimulus. If this motivated Higgins to engage in affiliative behaviour (or any other kind of behaviour that serves to preserve or intensify the other's joy), we could assert that he is behaving on the basis of an emotion that tracks the moral proposition "This creature's happiness is good." This tracking would obtain in virtue of the fact that Higgins' emotion has the adequate phenomenal character (both the descriptive and the motivational component of this proposition are present in Higgins' emotion), and there is a truth-preserving relation between Higgins' emotion and this proposition. This positive form of MME resembles our moral phenomenology, as it is close to what we experience when we are kind or encouraging to someone who is happy. It also entails a sensitivity to a morally relevant feature of situations (i.e. happiness—a positive or pleasant mental state), and can be both externally and internally evaluated as the morally right response for Higgins to have, given the situation. Positive MME thus qualifies as moral for the same reasons as its negative counterpart. And Higgins doesn't have to be capable of attributing mental states to others, either, for his emotional reaction need only be triggered by, and directed at, the superficial behavioural cues that accompany displays of joy in others.

If we attend to the normativity behind both the negative and the positive counterparts of MME, rather than to their phenomenal character, we would consider them both as *positive* or *good* moral emotions, insofar as internal and external evaluations would classify each of them, under normal circumstances, as the *morally right* thing to feel. Rowlands' framework allows us to construct further minimal accounts of other moral emotions that are also positive, in this normative sense, and that are not dependent upon the presence of emotional contagion. An example of this is the emotion of patience. Imagine that we witness old Higgins stoically putting up with the playful pestering of a young puppy. We may be inclined to describe Higgins' behaviour as 'patient.' As in the case of empathy, though, many may be hesitant to use such a label if we cannot confidently attribute mindreading capacities to him—if we cannot be sure that Higgins can understand that the puppy has no *intention* of doing him any harm, but merely wants to play. Using Rowlands' theory, we can construct a minimal account of patience that won't require the presence of mindreading capacities—what we might label *minimal moral patience*.

Suppose that Higgins possesses a 'moral module' that ensures that he inhibits any aggressive reaction he might have towards the behaviour of young puppies that he finds annoying. It's easy to imagine an evolutionary advantage that may come from a mechanism that ensures patience in response to the behaviour of young individuals, for then one's genetic offspring would have a greater chance of reaching a healthy maturity than those whose parents responded violently to their behaviour. Such a 'moral module' is, thus, not implausible. Suppose, then, that whenever a puppy bites him, Higgins inhibits his initial impulse to bite his aggressor back when he sees/hears/smells that it's just a puppy. Higgins would be behaving on the basis of an emotion that tracks the evaluative content "This puppy's behaviour should not be punished." As in the case of MME, the evaluative content of this emotion is not entertained by Higgins. Rather, when he inhibits his impulse to respond aggressively to the puppy's behaviour, both the descriptive and the motivational component of this moral proposition are present in this reaction. In responding this way, Higgins is further showing a sensitivity to the bad-making feature of situations that would be punishing the puppy for its behaviour. Higgins' reaction also resembles our phenomenology in at least some of the cases when we refrain from reacting angrily in response to a child's misbehaviour. By inhibiting his impulse to react aggressively, Higgins is contributing to the non-existence of something bad, which means his response can be externally evaluated as morally right. And because his reaction is motivated by his classification of the puppy as such, we can also evaluate it internally as the morally right thing to do, given the circumstances.

Rowlands' framework allows us, as well, to make sense of the idea that some animals' behaviour may sometimes be morally wrong, because it has been triggered by a moral emotion that is *evil* or *negative*, in the normative sense. We can thus avoid the charge of excessive romanticism that would come from painting a picture of nature as full of entirely benevolent nonhuman creatures. Let us imagine the case of Hans, who sometimes behaves on the basis of what we could call *minimal moral Schadenfreude*. In contrast to Higgins, Hans has a 'moral module' that makes him reliably undergo a reverse form of emotional contagion when seeing other individuals in distress. Like Higgins, Hans can discriminate the presence of distress behaviour in his immediate environment, and categorise it as such, but instead of making him feel an intentional form of distress, it makes him feel an intentional form of joy. Hans feels happy when he sees others in distress, where this happiness is triggered by, and directed at, the superficial behavioural cues that come with distress, and so doesn't require the presence of mindreading capacities, either. In addition to feeling happy when he sees others in distress, built into Hans' happiness is an urge to preserve the situation, or to somehow contribute to intensifying the target's distress behaviour.

We could describe Hans as behaving on the basis of an emotion that tracks the proposition "This creature's distress is good." The descriptive component of this proposition is present in Hans' enjoyment of others' distress (behaviour), and the motivational component is also there, for Hans behaves in a similar way as someone might if she believed this proposition to be true. If we assume, for the sake of the argument, that propositions that follow the form "X's distress is good" are always false, then we would say that Hans' emotion is always misguided, for he reliably experiences as good something that is not, in fact, good. Hans possesses a sensitivity to a morally-relevant feature of situations, but one that misfires. This misguided character of his emotion allows us to evaluate it, internally, as a morally wrong reason for him to contribute to others' distress. Hans behaves this way because he enjoys something that he *should not* enjoy. Given that behaviour motivated by this minimal form of Schadenfreude will tend to increase the amount of bad in the world, we can also evaluate it externally as morally wrong. While Hans' behaviour is misguided, it is not reprehensible (even if it should, arguably, be stopped), because he lacks an adequate understanding of the wrongness of what he is doing and so is not morally responsible for it. His lack of mindreading abilities, as well as his lack of moral judgement, precludes Hans from being a moral agent, but, insofar as some of his behaviour is motivated by his reliable tendency to experience others' distress as something enjoyable, he is a moral subject.

This brief analysis should be enough to show that Rowlands' framework can easily accommodate further moral emotions. This points to the idea that morality in a broader sense, and not just empathy, is independent of metacognition, mindreading, and moral judgement. Future research on animal morality could, I believe, greatly benefit from utilising this framework to construct de-intellectualised accounts of other moral emotions, such as compassion, guilt, indignation, resentment, etc., which could help us in our search for moral behaviour in nonhuman species.

6.3. SHOULD WE REVERSE THE QUESTION?

If moral emotions can indeed exist in the absence of mindreading capacities—which, as I've argued, is at least a theoretical possibility with some empirical plausibility—then the following question opens up: could minimal moral emotions have had a role to play in the evolutionary emergence of mindreading? As I intend to show now, there may be some *prima facie* reasons for thinking that minimal moral emotions could have facilitated the appearance of mindreading in phylogeny.

Kristin Andrews' account of the evolutionary emergence of mindreading abilities (Andrews 2009; 2012b; 2013; 2015) offers a promising framework for tackling this question. She criticises standard accounts of human folk psychology that presuppose a symmetry between the prediction and the explanation of behaviour, and assume that the attribution of propositional attitudes to others is the key ability that allows humans to both predict and explain the behaviour of others. Andrews argues that, contrary to these standard accounts, prediction and explanation in folk psychology are not symmetrical, for we can both predict behaviour that we cannot explain, and explain behaviour that we could not have predicted. Furthermore, she argues that these standard accounts are also wrong because humans rarely attribute propositional attitudes to others when it comes to predicting their behaviour. It is much more common for us to predict others' behaviour by using heuristics that do not require the intervention of our mindreading abilities: we predict from the situation, we generalise from our own experiences, we use stereotypes and personality traits to guess what others will do, etc. It is therefore unlikely that mindreading abilities emerged to help us predict the behaviour of others. On the other hand, she argues, it is more plausible to think that mindreading abilities developed to fulfill our need to explain behaviour. The key idea in her account is the following:

> It is odd behavior that we seek to explain; there is no need to explain normal behavior. Here comes the important step: to realize some behavior is worth explaining is to realize that it is not normal, and this requires an understanding of what is normal, or what individuals should be doing. It requires some understanding of social norms. (Andrews 2013, 119)

According to Andrews, then, the need to explain behaviour, which will to lead to the appearance of mindreading, can only emerge in the presence of anomalous behaviour. This presupposes the existence of norms, for only they can allow behaviour to appear as odd or anomalous. Thus, mindreading probably originated in a society that already had norms in place. Norms are conceived here as "an understanding of how we do things around here that doesn't rely on having personally accessible rules of action" (Andrews 2015, 60). This sort of understanding would make it possible for a member of a community to engage in behaviour that other members would read as anomalous, thus triggering in them a need to explain it that would eventually result in the development of mindreading capacities.

I suspect that there may be something missing in Andrews' account of the emergence of mindreading abilities. The mere presence of norms does not seem enough to secure the appearance of a need to explain anomalous behaviour. An important characteristic of societies with norms is the presence of sanctions. Indeed, this is what allows us to distinguish norms from mere statistical regularities of behaviour. While ants behave in fairly predictable ways, no one would describe ant colonies as being governed by social norms. And this is, at least partly, because ants do not sanction the behaviour of those that do not conform to the statistical regularity. The behaviour of ants is presumably the result of an innate fixed action pattern. Behavioural regularities that are the result of a social norm are acquired regularities, and sanctioning seems like a necessary mechanism for both the acquisition and the persistence of social norms. Within a very primitive society, the most pressing thing to do when faced with anomalous behaviour may thus be to sanction it in order to ensure the persistence of the norm. Therefore, the mere existence of norms might not guarantee the appearance of mindreading practices. We need something that will make the need to *explain* anomalous behaviour be stronger than the need to *sanction* it.

The missing link in this account may be minimal moral emotions. At least, there is a reason for thinking that MME could have a role to play here. This minimal moral emotion has the quality of making it hard for one to ignore others' distress. If an individual possesses MME, she will feel others' distress as unpleasant. At the same time, sanctions, especially when they take the form of physical punishment, can be a source of distress for those who are on the receiving end. In a community where there were norms but no mindreading capacities, one could subject others to punishment without a second thought. If, however, the members of such a community possessed MME, then punishing others for their anomalous behaviour becomes an unpleasant experience that they will prefer to avoid if possible. This means that they would have a *reason* to try and explain others' anomalous behaviour instead of merely sanctioning it. MME can, therefore, provide the missing link that ensures that anomalous behaviour is viewed as something that needs to be explained.

This idea does not have to be restricted to MME. Other moral emotions could, perhaps, be potential facilitators in this process, especially if they help foster a sense of mutural care and concern. To see this, let's look at an example that Andrews often uses:

> Consider our ancestors who have not yet discovered fire, but jointly hunt for meat.

> Fire is known, but only as a destructive force. After the hunt, I predict that you will eat your share of the meat. But that prediction is a failure, because you shove all the meat in the fire. What to do? We might all be outraged at your crazy behavior, knowing that fire is destructive and meat is only hard won. But if we consider you to be an in-group member, and want to keep you as part of the group, we can seek to explain. This might take the form of copying your behavior and eating the cooked meat. Learning that meat in fire tastes better than raw meat, we will understand something more about your motivation (Andrews 2015, 59)

What could move us to copy the behaviour of the offender in this example, instead of sanctioning her for seemingly destroying the meat? According to Andrews, the decisive factor would be our desire to keep her as a member of our group. What could have led us to this desire? It could, perhaps, be the result of a belief that one fares better in cooperation with others than alone. But this may be quite intellectually demanding. Another possibility, and perhaps a simpler one, is to postulate a mutual sense of care and concern facilitated by the existence, in this community, of minimal moral emotions. If we care for this individual, we will be more likely to want to keep her as part of our group. Moreover, if in this community we tend to be patient, trusting, and compassionate towards each other, we may be more inclined to first try to understand *why* she is throwing the meat in the fire, before reacting with punishment or ostracism.

Minimal moral emotions may well be the only mechanisms capable of generating this sort of mutual care and concern in the absence of mindreading capacities. In other texts, Andrews acknowledges that emotions have a role to play in her story, for they may help generate an affective tension in those who witness anomalous behaviour, thus giving rise to a need to explain it. The emotions she refers to—fear, puzzlement, curiosity, and disbelief (Andrews 2012, 120)—are not, however, moral emotions. Because they are not moral emotions, I suspect that they cannot do the work that Andrews wants them to do, unless we presuppose a context of mutual care and concern. If there is no mutual care and concern, there is nothing to stop fear, puzzlement, and disbelief from triggering sanctioning. As, *ex hypothesi*, Andrews' story takes place in a society with no mindreading, the presence of minimal moral emotions may be needed to ensure, or at least facilitate, the emergence of the sort of explanation-seeking practices that she postulates as key to the development of mindreading capacities.

Needless to say, the account I have offered of Andrews' theory is greatly simplified. Considering the role that minimal moral emotions might play here would require a much more detailed analysis of her claims. I have additionally presupposed much in this quick sketch, not least the plausibility of the story that Andrews presents, as well as the possibility of norms existing in such primitive communities. Further research would have to delve into these issues, as well as consider alternative accounts of the evolution of mindreading abilities. Still, this is enough to see that minimal moral emotions could be a promising factor to consider when attempting to explain the emergence of mindreading. My aim has simply been to offer some reasons for thinking that perhaps we should reverse our original question, and consider, not whether morality requires mindreading, but whether (the emergence of) mindreading requires morality.

6.4. ETHICAL CONSEQUENCES?

It is generally assumed that the kind of ethical treatment a certain being deserves depends on her particular characteristics. While this idea has been questioned by some (e.g. Crary 2010), it seems reasonable to suppose that not all beings can be subjected to the same kinds of harm, and that the ways in which a being can be harmed, in turn, shape the kinds of duties we might hold towards her. For instance, it makes no sense to say of a non-sentient being that she has a right not to be subjected to unnecessary pain. Taking this idea as our point of departure, we could ask whether any ethical consequences follow from a characterisation of (some) animals as moral subjects, that is, whether there are any specific kinds of harm that can affect moral subjects in virtue of the fact that they are such, and what rights or entitlements might follow from this.

If we focus on the case of MME, perhaps the most obvious ethical consequence that might follow from it emerges from the fact that those beings who possess this moral emotion, because it encompasses emotional contagion, will be affected by the distress of others. This consideration could be used to strengthen the ethical case against certain husbandry practices where animals routinely witness the pain and suffering of others, in those cases where the animals involved could reasonably be thought to possess MME (or emotional contagion, for that matter). While this line of reasoning should not be ignored, I believe that there are stronger ethical consequences that follow from the idea of MME in particular and moral subjecthood more generally; considerations that will help us move away from purely welfarist approaches to animal ethics. Emotional contagion may be an important source of suffering for certain animals, but its outcome is 'merely' distress, which does not represent a distinct type of harm. I believe that moral subjecthood brings to the table the possibility of a specific type of harm that might not be addressed by merely improving the welfare conditions faced by animals in captivity.

A way of specifying the ethical consequences that follow from moral subjecthood is to explore the capabilities approach to animal ethics, which was introduced by Martha Nussbaum (2004; 2007). Rather than focusing merely on pleasure and pain, the capabilities approach locates ethical significance in the existence of complex forms of life. The fact that complex forms of life generate wonder in us suggests that their flourishing is a good thing, and accordingly, the capabilities approach aims to "see each thing flourish as the sort of thing it is" (Nussbaum 2004, 306). A similar idea to this 'wonder' that Nussbaum refers to is present in Rowlands (2012b), who argues that moral subjects merit a very specific kind of attitude from us—what he calls 'moral respect' (Rowlands 2012b, chap. 10). Respect, according to Rowlands, is an attitude that differs from both *praise* and *admiration*. Praise is the sort of attitude one holds towards an individual who is responsible for what she does, and it can take both a moral and a non-moral form. One can (non-morally) praise an artist who has accomplished an inspiring work of art. In the realm of morality, only moral agents can be legitimate subjects of praise. Admiration, in turn, is directed not at agents, but at acts and their outcomes. One can admire the work of an artist or the selfless act of a moral agent. Acts themselves cannot, strictly speaking, be praised, only admired. As for respect, it is directed at actors, just like praise, and can also be used in a moral and in a non-moral sense. In contrast to praise, however, respect can be legitimately applied in the absence of responsibility. Respect is thus an attitude that can be held towards moral subjects, and should, in fact, be held when the moral subjects in question behave on the basis of 'good' moral emotions. In Rowlands' words:

> If we assume that the moral subject, in this instance, acts in a way that is good rather than evil, then we can say it is a good thing—morally speaking—that the world contains someone who acts in this way. … This is not moral praise, of the sort one might give to an agent, responsible for its actions. Nor is it moral admiration of the sort one might have for something that cannot act—an illuminating moral work of some sort. Rather, this is the sort of attitude one bears to something that can act, and acts for the good, but is not responsible for what it does. This is moral respect. If animals can, and sometimes do, act for moral reasons, then they are worthy objects of moral respect. (Rowlands 2012b, 253-4)

While the 'wonder' that Nussbaum refers to would apply to all complex forms of life, moral respect applies to a subset of these complex forms of life—those that can behave on the basis of good moral motivations. The fact that moral subjects are worthy of moral respect highlights the good that comes from them flourishing as the type of thing that they are. Moral subjects don't just fill us with aesthetic or intellectual wonder; they also make the world a better place, morally speaking. This points us in the direction of the idea that (some of) the moral motivations that grant moral subjecthood belong in the class of capabilities that Nussbaum considers “basic,” because they can be “evaluated as both good and central” (Nussbaum 2004, 309). In fact, those moral emotions that constitute positive forms of care for others could well be considered to be tacitly present in Nussbaum’s list of the central capabilities of animals—they could be easily filed under the categories of “emotion” or “affiliation,” both of which refer to (some) animals’ ability to “have attachments to others, to love and care for others” (Nussbaum 2004, 316).

By thinking of these minimal moral emotions as basic capabilities, we can see how they open the door to a distinct type of harm. Nussbaum’s capabilities approach locates ethical significance in the flourishing of the basic capabilities, which means that “it will also find harm in the thwarting or blighting of those capabilities” (Nussbaum 2004, 309). Individuals that are more complex than others will have more capabilities that can be thwarted, which results in the possibility of them being harmed in more and distinct ways. If minimal moral emotions are indeed basic capabilities, this means that the individuals who possess them are entitled to their flourishing. Whenever an animal is treated in a way that prevents these capabilities from flourishing—because, say, it is kept in isolation from others, or subjected to a relevant form of conditioning or training—, the animal is being harmed. This, in turn, would be a distinct type of harm that would occur regardless of whether the animal ever actually suffers as a result of this treatment, and of whether she is capable of understanding what is missing in her life.

The characterisation of certain animals as moral subjects may, therefore, have important ethical consequences. Not only could it be used to reinforce the ethical case against certain practices involving animals; it may even give rise to certain specific entitlements or rights that these animals would otherwise lack. Future investigations could be directed at evaluating our treatment of animals in light of these considerations. I believe—and this is an entirely personal claim—that this should be the ultimate aim of all research into the minds of animals.

REFERENCES

"15 Years Ago Today: Gorilla Rescues Boy Who Fell In Ape Pit." 2011. August 16. http://chicago.cbslocal.com/2011/08/16/15-years-ago-today-gorilla-rescues-boy-who-fell-in-ape-pit/.

Abbate, Cheryl. 2014. "Nonhuman Animals: Not Necessarily Saints or Sinners." *Between the Species* 17 (1). doi: 10.15368/bts.2014v17n1.4

Albuquerque, Natalia, Kun Guo, Anna Wilkinson, Carine Savalli, Emma Otta, and Daniel Mills. 2016. "Dogs Recognize Dog and Human Emotions." *Biology Letters* 12 (1): 20150883. doi:10.1098/rsbl.2015.0883.

Alexander, Larry, and Michael Moore. 2012. "Deontological Ethics." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2012. http://plato.stanford.edu/archives/win2012/entries/ethics-deontological/.

Allen, Colin. 1999. "Animal Concepts Revisited: The Use of Self-Monitoring as an Empirical Approach." *Erkenntnis* 51 (1): 537–44. doi:10.1023/A:1005545425672.

Anderson, James R., Masako Myowa–Yamakoshi, and Tetsuro Matsuzawa. 2004. "Contagious Yawning in Chimpanzees." *Proceedings of the Royal Society of London B: Biological Sciences* 271 (Suppl 6): S468–70. doi:10.1098/rsbl.2004.0224.

Andrews, Kristin. 2002. "Interpreting Autism: A Critique of Davidson on Thought and Language." *Philosophical Psychology* 15 (3): 317–32.

———. 2005. "Chimpanzee Theory of Mind: Looking in All the Wrong Places?" *Mind & Language* 20 (5): 521–36. doi:10.1111/j.0268-1064.2005.00298.x.

———. 2009. "Understanding Norms without a Theory of Mind." *Inquiry* 52 (5): 433–48. doi:10.1080/00201740903302584.

———. 2012a. Review of *Mindreading Animals: The Debate over What Animals Know about Other Minds*, by Robert W. Lurz. *Notre Dame Philosophical Reviews*, March.

———. 2012b. *Do Apes Read Minds?: Toward a New Folk Psychology*. Cambridge, MA: MIT Press.

———. 2013. "Great Ape Mindreading: What's at Stake?" In *The Politics of Species: Reshaping our Relationships with Other Animals*, edited by Raymond Corbey and Annette Lanjouw, 115–25. Cambridge: Cambridge University Press.

———. 2015. "The Folk Psychological Spiral: Explanation, Regulation, and Language." *The Southern Journal of Philosophy* 53 (September): 50–67. doi:10.1111/sjp.12121.

Andrews, Kristin, and Lori Gruen. 2014. "Empathy in Other Apes." In *Empathy and Morality*, edited by Heidi Maibom, 193–209. Oxford: Oxford University Press.

Andrews, Kristin, and Brian Huss. 2014. "Anthropomorphism, Anthropectomy, and the Null Hypothesis." *Biology and Philosophy* 29 (5): 711–29. doi:10.1007/s10539-014-9442-2.

Anil, M. Haluk, John Preston, Justin L. McKinstry, Richard G. Rodway, and Steven N. Brown. 1996. "An Assessment of Stress Caused in Sheep by Watching Slaughter of Other Sheep." *Animal Welfare* 5 (4): 435–41.

Atsak, Piray, Marie Orre, Petra Bakker, Leonardo Cerliani, Benno Roozendaal, Valeria Gazzola, Marta Moita, and Christian Keysers. 2011. "Experience Modulates Vicarious Freezing in Rats: A Model for Empathy." *PLoS ONE* 6 (7): e21855. doi:10.1371/journal.pone.0021855.

Aureli, Filippo, Roberto Cozzolino, Carla Cordischi, and Stefano Scucchi. 1992. "Kin-Oriented Redirection among Japanese Macaques: An Expression of a Revenge System?" *Animal Behaviour* 44, Part 2 (August): 283–91. doi:10.1016/0003-3472(92)90034-7.

Baillargeon, Renée, Rose M. Scott, and Zijing He. 2010. "False-Belief Understanding in Infants." *Trends in Cognitive Sciences* 14 (3): 110–18. doi:10.1016/j.tics.2009.12.006.

Baird, Jodie A., and Janet Wilde Astington. 2004. "The Role of Mental State Understanding in the Development of Moral Cognition and Moral Action." *New Directions for Child and Adolescent Development* 2004 (103): 37–49. doi:10.1002/cd.96.

Bard, Kim A. 2006. "Neonatal Imitation in Chimpanzees (Pan Troglodytes) Tested with Two Paradigms." *Animal Cognition* 10 (2): 233–42. doi:10.1007/s10071-006-0062-3.

Baron-Cohen, Simon, Alan M. Leslie, and Uta Frith. 1985. "Does the Autistic Child Have a 'Theory of Mind'?" *Cognition* 21: 37–46.

Bartal, Inbal Ben-Ami, Jean Decety, and Peggy Mason. 2011. "Empathy and Pro-Social Behavior in Rats." *Science* 334 (6061): 1427–30. doi:10.1126/science.1210789.

Bates, Lucy A., Richard Byrne, Phyllis C. Lee, Norah Njiraini, Joyce H. Poole, Katito Sayialel, Soila Sayialel, and Cynthia J. Moss. 2008. "Do Elephants Show Empathy?" *Journal of Consciousness Studies* 15 (10-11): 204–25.

Bates, Lucy A., and Richard W. Byrne. 2007. "Creative or Created: Using Anecdotes to Investigate Animal Cognition." *Methods (San Diego, Calif.)* 42 (1): 12–21. doi:10.1016/j.ymeth.2006.11.006.

Batson, C. Daniel. 1991. *The Altruism Question: Toward a Social-Psychological Answer*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

———. 2009. "These Things Called Empathy: Eight Related but Distinct Phenomena." In *The Social Neuroscience of Empathy*, edited by Jean Decety and William Ickes, 3–15. Cambridge, MA: MIT Press.

Bekoff, Marc. 1977. "Social Communication in Canids: Evidence for the Evolution of a

Stereotyped Mammalian Display." *Science* 197 (4308): 1097–99. doi:10.1126/science.197.4308.1097.

———. 2001. "Social Play Behaviour: Cooperation, Fairness, Trust, and the Evolution of Morality." *Journal of Consciousness Studies* 8: 81–90.

———. 2004. "Wild Justice and Fair Play: Cooperation, Forgiveness, and Morality in Animals." *Biology and Philosophy* 19 (4): 489–520. doi:10.1007/sBIPH-004-0539-x.

Bekoff, Marc, and Jessica Pierce. 2009. *Wild Justice: The Moral Lives of Animals*. Chicago, IL: University Of Chicago Press.

Bello, Paul, and Selmer Bringsjord. 2013. "On How to Build a Moral Machine." *Topoi* 32 (2): 251–66. doi:10.1007/s11245-012-9129-8.

Berntson, Gary G., Sarah T. Boysen, Harold R. Bauer, and Michael S. Torello. 1989. "Conspecific Screams and Laughter: Cardiac and Behavioral Reactions of Infant Chimpanzees." *Developmental Psychobiology* 22 (8): 771–87. doi:10.1002/dev.420220803.

Blackburn, Simon. 1993. *Essays in Quasi-Realism*. New York, NY: Oxford University Press.

Blair, James R., and Karina S. Perschardt. 2002. "Empathy: A Unitary Circuit or a Set of Dissociable Neuro-Cognitive Systems?" *Behavioral and Brain Sciences* 25 (1): 27–28.

Boesch, Christophe. 2002. "Cooperative Hunting Roles among Taï Chimpanzees." *Human Nature* 13 (1): 27–46. doi:10.1007/s12110-002-1013-6.

"Boy Hails Cat Tara That Saved Him from Dog: 'She's a Hero!'" 2014. *The Independent*. May 15. http://www.independent.co.uk/news/weird-news/boy-hails-cat-tara-that-saved-him-from-dog-shes-a-hero-9375674.html.

Bradie, Michael. 1994. *The Secret Chain: Evolution and Ethics*. New York, NY: State University of New York Press.

Brosnan, Sarah F. 2006. "Nonhuman Species' Reactions to Inequity and Their Implications for Fairness." *Social Justice Research* 19 (2): 153–85. doi:10.1007/PL00022136.

Brosnan, Sarah F., and Frans B.M. de Waal. 2003. "Monkeys Reject Unequal Pay." *Nature* 425 (6955): 297–99. doi:10.1038/nature01963.

Brosnan, Sarah F., Hillary C. Schiff, and Frans B.M. de Waal. 2005. "Tolerance for Inequity May Increase with Social Closeness in Chimpanzees." *Proceedings of the Royal Society B: Biological Sciences* 272 (1560): 253–58. doi:10.1098/rspb.2004.2947.

Brosnan, Sarah F., Catherine Talbot, Megan Ahlgren, Susan P. Lambeth, and Steven J. Schapiro. 2010. "Mechanisms Underlying Responses to Inequitable Outcomes in Chimpanzees, Pan Troglodytes." *Animal Behaviour* 79 (6): 1229–37. doi:10.1016/j.anbehav.2010.02.019.

Buckner, Cameron. 2013. "Morgan's Canon, Meet Hume's Dictum: Avoiding Anthropofabulation in Cross-Species Comparisons." *Biology and Philosophy* 28 (5): 853–71.

———. 2014. "The Semantic Problem(s) with Research on Animal Mindreading." *Mind and Language* 29 (5): 566–89.

Bugnyar, Thomas, and Bernd Heinrich. 2005. "Ravens, Corvus Corax, Differentiate between Knowledgeable and Ignorant Competitors." *Proceedings of the Royal Society B: Biological Sciences* 272 (1573): 1641–46. doi:10.1098/rspb.2005.3144.

Bugnyar, Thomas, and Kurt Kotrschal. 2002. "Observational Learning and the Raiding of Food Caches in Ravens, Corvus Corax: Is It 'Tactical' Deception?" *Animal Behaviour* 64 (2): 185–95. doi:10.1006/anbe.2002.3056.

Bugnyar, Thomas, Stephan A. Reber, and Cameron Buckner. 2016. "Ravens Attribute Visual Access to Unseen Competitors." *Nature Communications* 7 (February): 10506. doi:10.1038/ncomms10506.

Burge, Tyler. 1978. "Concept of Mind in Primates?" *Behavioral and Brain Sciences* 1 (4): 560–61.

Burkart, Judith M., Ernst Fehr, Charles Efferson, and Carel P. van Schaik. 2007. "Other-Regarding Preferences in a Non-Human Primate: Common Marmosets Provision Food Altruistically." *Proceedings of the National Academy of Sciences* 104 (50): 19762–66. doi:10.1073/pnas.0710310104.

Burkett, James P., Elissar Andari, Zachary V. Johnson, Daniel C. Curry, Frans B.M. de Waal, and Larry J. Young. 2016. "Oxytocin-Dependent Consolation Behavior in Rodents." *Science* 351 (6271): 375–78. doi:10.1126/science.aac4785.

Buttelmann, David, Josep Call, and Michael Tomasello. 2009. "Do Great Apes Use Emotional Expressions to Infer Desires?" *Developmental Science* 12 (5): 688–98. doi:10.1111/j.1467-7687.2008.00802.x.

Buttelmann, David, and Michael Tomasello. 2013. "Can Domestic Dogs (Canis Familiaris) Use Referential Emotional Expressions to Locate Hidden Food?" *Animal Cognition* 16 (1): 137–45. doi:10.1007/s10071-012-0560-4.

Butterfill, Stephen A., and Ian A. Apperly. 2013. "How to Construct a Minimal Theory of Mind." *Mind & Language* 28 (5): 606–37. doi:10.1111/mila.12036.

Byrne, Richard W., and Lucy A. Bates. 2011. "Cognition in the Wild: Exploring Animal Minds with Observational Evidence." *Biology Letters* 7 (4): 619–22. doi:10.1098/rsbl.2011.0352.

Bzdok, Danilo, Leonhard Schilbach, Kai Vogeley, Karla Schneider, Angela R. Laird, Robert

Langner, and Simon B. Eickhoff. 2012. "Parsing the Neural Correlates of Moral Cognition: ALE Meta-Analysis on Morality, Theory of Mind, and Empathy." *Brain Structure and Function* 217 (4): 783–96. doi:10.1007/s00429-012-0380-y.

Call, Josep, Brian Hare, Malinda Carpenter, and Michael Tomasello. 2004. "'Unwilling' versus 'unable': Chimpanzees' Understanding of Human Intentional Action." *Developmental Science* 7 (4): 488–98. doi:10.1111/j.1467-7687.2004.00368.x.

Call, Josep, Brian Hare, and Michael Tomasello. 1998. "Chimpanzee Gaze Following in an Object-Choice Task." *Animal Cognition* 1 (2): 89–99. doi:10.1007/s100710050013.

Call, Josep, and Michael Tomasello. 1999. "A Nonverbal False Belief Task: The Performance of Children and Great Apes." *Child Development* 70 (2): 381–95.

———. 2008. "Does the Chimpanzee Have a Theory of Mind? 30 Years Later." *Trends in Cognitive Sciences* 12 (5): 187–92. doi:10.1016/j.tics.2008.02.010.

Campbell, Matthew W., J. Devyn Carter, Darby Proctor, Michelle L. Eisenberg, and Frans B.M. de Waal. 2009. "Computer Animations Stimulate Contagious Yawning in Chimpanzees." *Proceedings of the Royal Society of London B: Biological Sciences*, 276 (1676): 4255–59. doi:10.1098/rspb.2009.1087.

Campbell, Matthew W., and Frans B.M. de Waal. 2014. "Chimpanzees Empathize with Group Mates and Humans, but Not with Baboons or Unfamiliar Chimpanzees." *Proceedings of the Royal Society of London B: Biological Sciences* 281 (1782): 20140013. doi:10.1098/rspb.2014.0013.

Campos Serena, Olga. 2012. "La Relevancia de Ser Un Sujeto Moral. Comentario Al Artículo de Mark Rowlands '¿Pueden Los Animales Ser Morales?'" *Dilemata* 0 (9): 75–82.

Carr, Laurie, Marco Iacoboni, Marie-Charlotte Dubeau, John C. Mazziotta, and Gian Luigi Lenzi. 2003. "Neural Mechanisms of Empathy in Humans: A Relay from Neural Systems for Imitation to Limbic Areas." *Proceedings of the National Academy of Sciences* 100 (9): 5497–5502. doi:10.1073/pnas.0935845100.

Carruthers, Peter. 2009. "How We Know Our Own Minds: The Relationship between Mindreading and Metacognition." *The Behavioral and Brain Sciences* 32 (2): 121–38; discussion 138–82. doi:10.1017/S0140525X09000545.

"Center For Great Apes :: Knuckles." 2016. Accessed March 14. http://www.centerforgreatapes.org/meet-apes/chimpanzees/knuckles/.

Chalmeau, Raphael. 1994. "Do Chimpanzees Cooperate in a Learning Task?" *Primates* 35 (3): 385–92. doi:10.1007/BF02382735.

Chater, Nick, and Cecilia M. Heyes. 1994. "Animal Concepts: Content and Discontent." *Mind and Language* 9 (3): 209–46.

Church, Russell M. 1959. “Emotional Reactions of Rats to the Pain of Others.” *Journal of Comparative and Physiological Psychology* 52 (2): 132–34. doi: 10.1037/h0043531.

Clay, Zanna, and Frans B.M. de Waal. 2013. “Bonobos Respond to Distress in Others: Consolation across the Age Spectrum.” *PLoS ONE* 8 (1): e55206. doi:10.1371/journal.pone.0055206.

Colditz, Ian G., David R. Paull, and Caroline Lee. 2012. “Social Transmission of Physiological and Behavioural Responses to Castration in Suckling Merino Lambs.” *Applied Animal Behaviour Science* 136 (2): 136–45. doi:10.1016/j.applanim.2011.12.011.

Connor, Richard, and Kenneth Norris. 1982. “Are Dolphins Reciprocal Altruists?” *The American Naturalist* 119 (3): 358–74.

Cools, Annemieke K.A., Alain J.-M. Van Hout, and Mark H. J. Nelissen. 2008. “Canine Reconciliation and Third-Party-Initiated Postconflict Affiliation: Do Peacemaking Social Mechanisms in Dogs Rival Those of Higher Primates?” *Ethology* 114 (1): 53–63. doi:10.1111/j.1439-0310.2007.01443.x.

Cordoni, Giada, Elisabetta Palagi, and Silvana Borgognini Tarli. 2006. “Reconciliation and Consolation in Captive Western Gorillas.” *International Journal of Primatology* 27 (5): 1365–82. doi:10.1007/s10764-006-9078-4.

Cova, Florian. 2013. “Two Kinds of Moral Competence: Moral Agent, Moral Judge.” In *What Makes Us Moral? On the Capacities and Conditions for Being Moral*, edited by Bert Musschenga and Anton van Harskamp, 117–30. Library of Ethics and Applied Philosophy 31. Dordrecht: Springer Netherlands.

Cozzi, Alessandro, Claudio Sighieri, Angelo Gazzano, Christine J. Nicol, and Paolo Baragli. 2010. “Post-Conflict Friendly Reunion in a Permanent Group of Horses (Equus Caballus).” *Behavioural Processes* 85 (2): 185–90. doi:10.1016/j.beproc.2010.07.007.

Crary, Alice. 2010. “Minding What Already Matters: A Critique of Moral Individualism.” *Philosophical Topics* 38 (1): 17–49. doi:10.5840/philtopics20103812.

Crawford, Meredith P. 1937. *The Cooperative Solving of Problems by Young Chimpanzees.* Baltimore, MD: Johns Hopkins Press.

Creel, Scott, and Nancy Marusha Creel. 1995. “Communal Hunting and Pack Size in African Wild Dogs, Lycaon Pictus.” *Animal Behaviour* 50 (5): 1325–39. doi:10.1016/0003-3472(95)80048-4.

Cronin, Katherine A., Aimee V. Kurian, and Charles T. Snowdon. 2005. “Cooperative Problem Solving in a Cooperatively Breeding Primate (Saguinus Oedipus).” *Animal Behaviour* 69 (1): 133–42. doi:10.1016/j.anbehav.2004.02.024.

Cronin, Katherine A., Kori K. E. Schroeder, and Charles T. Snowdon. 2010. “Prosocial

Behaviour Emerges Independent of Reciprocity in Cottontop Tamarins." *Proceedings of the Royal Society of London B: Biological Sciences* 277 (1701): 3845–51. doi:10.1098/rspb.2010.0879.

Cronin, Katherine A., and Charles T. Snowdon. 2008. "The Effects of Unequal Reward Distributions on Cooperative Problem Solving by Cottontop Tamarins, Saguinus Oedipus." *Animal Behaviour* 75 (1): 245–57. doi:10.1016/j.anbehav.2007.04.032.

Cushman, Fiery. 2008. "Crime and Punishment: Distinguishing the Roles of Causal and Intentional Analyses in Moral Judgment." *Cognition* 108 (2): 353–80. doi:10.1016/j.cognition.2008.03.006.

Cushman, Fiery, Rachel Sheketoff, Sophie Wharton, and Susan Carey. 2013. "The Development of Intent-Based Moral Judgment." *Cognition* 127: 6–21.

Custance, Deborah, and Jennifer Mayer. 2012. "Empathic-like Responding by Domestic Dogs (Canis Familiaris) to Distress in Humans: An Exploratory Study." *Animal Cognition* 15 (5): 851–59.

Dally, Joanna M., Nathan J. Emery, and Nicola S. Clayton. 2004. "Cache Protection Strategies by Western Scrub-Jays (Aphelocoma Californica): Hiding Food in the Shade." *Proceedings of the Royal Society of London B: Biological Sciences* 271 Suppl. 6 (December): S387–90. doi:10.1098/rsbl.2004.0190.

———. 2005. "Cache Protection Strategies by Western Scrub-Jays, Aphelocoma Californica: Implications for Social Cognition." *Animal Behaviour* 70 (6): 1251–63. doi:10.1016/j.anbehav.2005.02.009.

———. 2006. "Food-Caching Western Scrub-Jays Keep Track of Who Was Watching When." *Science* 312 (5780): 1662–65. doi:10.1126/science.1126539.

Damasio, Antonio R. 1994. *Descartes' Error: Emotion, Reason, and the Human Brain*. First Edition edition. New York, NY: G.P. Putnam's Sons.

Dancy, Jonathan. 2000. *Practical Reality*. Oxford; New York, NY: Oxford University Press.

Daniel, William J. 1942. "Cooperative Problem Solving in Rats." *Journal of Comparative Psychology* 34 (3): 361–68. doi:10.1037/h0062264.

Dapretto, Mirella, Mari S. Davies, Jennifer H. Pfeifer, Ashley A. Scott, Marian Sigman, Susan Y. Bookheimer, and Marco Iacoboni. 2006. "Understanding Emotions in Others: Mirror Neuron Dysfunction in Children with Autism Spectrum Disorders." *Nature Neuroscience* 9 (1): 28–30. doi:10.1038/nn1611.

Darley, John M., Ellen Chereskin Klosson, and Mark P. Zanna. 1978. "Intentions and Their Contexts in the Moral Judgments of Children and Adults." *Child Development* 49 (1): 66–74. doi:10.2307/1128594.

Davidson, Donald. 1982. “Rational Animals.” *Dialectica* 36 (4): 317–27. doi:10.1111/j.1746-8361.1982.tb01546.x.

Davila Ross, Marina, Susanne Menzler, and Elke Zimmermann. 2008. “Rapid Facial Mimicry in Orangutan Play.” *Biology Letters* 4 (1): 27–30. doi:10.1098/rsbl.2007.0535.

de Boer, Jelle. 2011. “Moral Ape Philosophy.” *Biology & Philosophy* 26 (6): 891–904. doi:10.1007/s10539-011-9283-1.

Decety, Jean, and Philip L. Jackson. 2004. “The Functional Architecture of Human Empathy.” *Behavioral and Cognitive Neuroscience Reviews* 3 (2): 71–100. doi:10.1177/1534582304267187.

Decety, Jean, and Claus Lamm. 2009. “Empathy versus Personal Distress: Recent Evidence from Social Neuroscience.” In *The Social Neuroscience of Empathy*, edited by Jean Decety and William Ickes, 199–214. Cambridge, MA: MIT Press.

DeGrazia, David. 1996. *Taking Animals Seriously: Mental Life and Moral Status*. Cambridge: Cambridge University Press.

de Lora, Pablo. 2012. “¿Puede Gracia Ser Franciscana? Un Comentario a Mark Rowlands: ‘¿Pueden Los Animales Ser Morales?’” *Dilemata* 0 (9): 33–39.

Dennett, Daniel C. 1978. “Beliefs about Beliefs.” *Behavioral and Brain Sciences* 1 (4): 568–70.

de Vignemont, Frédérique, and Pierre Jacob. 2012. “What Is It like to Feel Another’s Pain?” *Philosophy of Science* 79 (2): 295–316. doi:10.1086/664742.

de Vignemont, Frédérique, and Tania Singer. 2006. “The Empathic Brain: How, When and Why?” *Trends in Cognitive Sciences* 10 (10): 435–41. doi:10.1016/j.tics.2006.08.008.

de Waal, Frans B.M. 1996. *Good Natured: The Origins of Right and Wrong in Humans and Other Animals*. Cambridge, MA: Harvard University Press.

———. 1997. *Bonobo: The Forgotten Ape*. 1st ed. Berkeley, CA: University of California Press.

———. 2006. *Primates and Philosophers: How Morality Evolved*. Edited by Josiah Ober and Stephen Macedo. Princeton, NJ: Princeton University Press.

———. 2008. “Putting the Altruism Back into Altruism: The Evolution of Empathy.” *Annual Review of Psychology* 59: 279–300. doi:10.1146/annurev.psych.59.103006.093625.

de Waal, Frans B.M., and Michelle L. Berger. 2000. “Payment for Labour in Monkeys.” *Nature* 404 (6778): 563–563. doi:10.1038/35007138.

de Waal, Frans B.M., and Lesleigh M. Luttrell. 1988. “Mechanisms of Social Reciprocity in Three Primate Species: Symmetrical Relationship Characteristics or Cognition?” *Ethology and Sociobiology* 9 (2): 101–18. doi:10.1016/0162-3095(88)90016-7.

de Waal, Frans B.M., and Angeline van Roosmalen. 1979. “Reconciliation and Consolation among Chimpanzees." *Behavioral Ecology and Sociobiology* 5 (1): 55–66. doi:10.1007/BF00302695.

Diéguez, Antonio. 2011. "Conceptual Thinking in Animals: Some Reflections on Language, Concepts, and Mind." In *Darwin's Evolving Legacy*, edited by Jorge Martínez-Contreras and Aura Ponce de León, 383–95. Mexico: Siglo XXI and Universidad Veracruzana.

di Pellegrino, Giuseppe, Luciano Fadiga, Leonardo Fogassi, Vittorio Gallese, and Giacomo Rizzolatti. 1992. "Understanding Motor Events: A Neurophysiological Study." *Experimental Brain Research* 91 (1): 176–80.

Dixon, Beth A. 2008. *Animals, Emotion & Morality: Marking the Boundary*. Amherst, NY: Prometheus Books.

Doris, John, and Stephen Stich. 2014. "Moral Psychology: Empirical Approaches." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2014. http://plato.stanford.edu/archives/spr2014/entries/moral-psych-emp/.

Douglas-Hamilton, Iain, Shivani Bhalla, George Wittemyer, and Fritz Vollrath. 2006. "Behavioural Reactions of Elephants towards a Dying and Deceased Matriarch." *Applied Animal Behaviour Science* 100 (1–2): 87–102. doi:10.1016/j.applanim.2006.04.014.

Dretske, Fred. 1969. *Seeing And Knowing*. Chicago, IL: University Of Chicago Press.

Driver, Julia. 2006. *Uneasy Virtue*. First paperback version. New York, NY: Cambridge University Press.

Dunn, Judy, Alexandra L. Cutting, and Helen Demetriou. 2000. "Moral Sensibility, Understanding Others, and Children's Friendship Interactions in the Preschool Period." *British Journal of Developmental Psychology* 18 (2): 159–77. doi:10.1348/026151000165625.

Edgar, Joanne L., John C. Lowe, Elizabeth S. Paul, and Christine J. Nicol. 2011. "Avian Maternal Response to Chick Distress." *Proceedings of the Royal Society of London B: Biological Sciences* 278 (1721): 3129–34. doi:10.1098/rspb.2010.2701.

Edgar, Joanne L., Elizabeth S. Paul, Lauren Harris, Sarah Penturn, and Christine J. Nicol. 2012. "No Evidence for Emotional Empathy in Chickens Observing Familiar Adult Conspecifics." *PLoS ONE* 7 (2): e31542. doi:10.1371/journal.pone.0031542.

Eisenberg, Nancy, and Natalie Eggum. 2009. "Empathic Responding: Sympathy and Personal Distress." In *The Social Neuroscience of Empathy*, edited by Jean Decety and William Ickes, 71–84. Cambridge, MA: MIT Press.

Emery, Nathan J., and Nicola S. Clayton. 2001. "Effects of Experience and Social Context on

Prospective Caching Strategies by Scrub Jays." *Nature* 414 (6862): 443–46. doi:10.1038/35106560.

Emery, Nathan J., Joanna M. Dally, and Nicola S. Clayton. 2004. "Western Scrub-Jays (Aphelocoma Californica) Use Cognitive Strategies to Protect Their Caches from Thieving Conspecifics." *Animal Cognition* 7 (1): 37–43. doi:10.1007/s10071-003-0178-7.

Eslinger, Paul J. 1998. "Neurological and Neuropsychological Bases of Empathy." *European Neurology* 39 (4): 193–99.

Evans, Becky. 2013. "Disabled Killer Whale with Missing Fins Survives with the Help of Family Who Hunt for Its Food." *Mail Online*. Accessed May 25. http://www.dailymail.co.uk/news/article-2326868/Disabled-killer-whale-missing-fins-survives-help-family-hunt-food.html.

Evans, Valerie E., and William G. Braud. 1969. "Avoidance of a Distressed Conspecific." *Psychonomic Science* 15 (3): 166. doi:10.3758/BF03336261.

Experiment 3 Chimpanzee Helps in Experimental Condition. 2007. Video S5 from Warneken F., Hare B., Melis A., Hanus D., Tomasello M. "Spontaneous Altruism by Chimpanzees and Young Children". PLOS Biology. doi:10.1371/journal.pbio.0050184. PMID 17594177. PMC: 1896184. https://commons.wikimedia.org/wiki/File:Spontaneous-Altruism-by-Chimpanzees-and-Young-Children-pbio.0050184.sv005.ogv.

Fadiga, Luciano, Leonardo Fogassi, Giovanni Pavesi, and Giacomo Rizzolatti. 1995. "Motor Facilitation during Action Observation: A Magnetic Stimulation Study." *Journal of Neurophysiology* 73 (6): 2608–11.

Flombaum, Jonathan I., and Laurie R. Santos. 2005. "Rhesus Monkeys Attribute Perceptions to Others." *Current Biology: CB* 15 (5): 447–52. doi:10.1016/j.cub.2004.12.076.

Francescotti, Robert. 2007. "Animal Minds and Animal Ethics: An Introduction." *The Journal of Ethics* 11 (3): 239–52.

Fraser, Orlaith N., and Thomas Bugnyar. 2010. "Do Ravens Show Consolation? Responses to Distressed Others." *PLoS ONE* 5 (5): e10605. doi:10.1371/journal.pone.0010605.

Fraser, Orlaith N., Daniel Stahl, and Filippo Aureli. 2008. "Stress Reduction through Consolation in Chimpanzees." *Proceedings of the National Academy of Sciences* 105 (25): 8557–62. doi:10.1073/pnas.0804141105.

Gallagher, Shaun. 2008. "Direct Perception in the Intersubjective Context." *Consciousness and Cognition* 17: 535–43. doi:doi:10.1016/j.concog.2008.03.003.

———. 2012. "Empathy, Simulation, and Narrative." *Science in Context* 25 (03): 355–81. doi:10.1017/S0269889712000117.

Gallagher, Shaun, and Daniel J. Povinelli. 2012. "Enactive and Behavioral Abstraction Accounts of Social Understanding in Chimpanzees, Infants, and Adults." *Review of Philosophy and Psychology* 3 (1): 145–69.

Gallagher, Shaun, and Dan Zahavi. 2012. *The Phenomenological Mind*. 2nd edition. London; New York, NY: Routledge.

Gallese, Vittorio. 2003. "The Roots of Empathy: The Shared Manifold Hypothesis and the Neural Basis of Intersubjectivity." *Psychopathology* 36 (4): 171–80. doi:10.1159/000072786.

Gallese, Vittorio, and Alvin I. Goldman. 1998. "Mirror Neurons and the Simulation Theory of Mind-Reading." *Trends in Cognitive Sciences* 2 (12): 493–501.

Gallup, Andrew C., Lexington Swartwood, Janine Militello, and Serena Sackett. 2015. "Experimental Evidence of Contagious Yawning in Budgerigars (Melopsittacus Undulatus)." *Animal Cognition*, May, 1–8. doi:10.1007/s10071-015-0873-1.

Gallup Jr., Gordon G. 1982. "Self-Awareness and the Emergence of Mind in Primates." *American Journal of Primatology* 2 (3): 237–48. doi:10.1002/ajp.1350020302.

———. 1998. "Can Animals Empathize? Yes." *Scientific American* 9: 66–71.

Gallup Jr., Gordon G., and Steven M. Platek. 2002. "Cognitive Empathy Presupposes Self-Awareness: Evidence From Phylogeny, Ontogeny, Neuropsychology, and Mental Illness." *Behavioral and Brain Sciences* 25 (1): 36–37.

Galvan, Moriah, and Jennifer Vonk. 2016. "Man's Other Best Friend: Domestic Cats (F. Silvestris Catus) and Their Discrimination of Human Emotion Cues." *Animal Cognition* 19 (1): 193–205. doi:10.1007/s10071-015-0927-4.

Glock, Hans-Johann. 2000. "Animals, Thoughts And Concepts." *Synthese* 123 (1): 35–64. doi:10.1023/A:1005295521736.

Goldman, Alvin I. 1992. "Empathy, Mind, and Morals." *Proceedings and Addresses of the American Philosophical Association* 66 (3): 17–41. doi:10.2307/3130659.

———. 2006. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. 1st ed. New York, NY: Oxford University Press.

Gollwitzer, Mario, and Markus Denzler. 2009. "What Makes Revenge Sweet: Seeing the Offender Suffer or Delivering a Message?" *Journal of Experimental Social Psychology* 45: 840–44.

Gomila, Antoni. 2012. "A Naturalistic Defense of 'Human Only' Moral Subjects." *Dilemata* 0 (9): 69–73.

González-Torre, Ángel Pelayo. 2012. "Que La Moral Cuide de Los Animales. Comentario Al Artículo de Mark Rowlands '¿Pueden Los Animales Ser Morales?'" *Dilemata* 0 (9):

135–40.

Goodall, Jane. 1990. *Through a Window: My Thirty Years with the Chimpanzees of Gombe*. Boston, MA: Houghton Mifflin.

Gordon, Robert M. 1986. “Folk Psychology as Simulation.” *Mind and Language* 1 (2): 158–71.

———. 2009. “Folk Psychology as Mental Simulation.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2009. http://plato.stanford.edu/archives/fall2009/entries/folkpsych-simulation/.

Gray, Kurt, Liane Young, and Adam Waytz. 2012. “Mind Perception Is the Essence of Morality.” *Psychological Inquiry* 23 (2): 101–24. doi:10.1080/1047840X.2012.651387.

Greene, James Thomas. 1969. “Altruistic Behavior in the Albino Rat.” *Psychonomic Science* 14 (1): 47–48. doi:10.3758/BF03336420.

Gruen, Lori. 2015. *Entangled Empathy: An Alternative Ethic for Our Relationships with Animals*. New York, NY: Lantern Books.

Guglielmo, Steve, Andrew E. Monroe, and Bertram F. Malle. 2009. “At the Heart of Morality Lies Folk Psychology.” *Inquiry* 52 (5): 449–66. doi:10.1080/00201740903302600.

Hampton, Robert R. 2009. “Multiple Demonstrations of Metacognition in Nonhumans: Converging Evidence or Multiple Mechanisms?” *Comparative Cognition & Behavior Reviews* 4 (January): 17–28.

Hare, Brian, Josep Call, and Michael Tomasello. 2001. “Do Chimpanzees Know What Conspecifics Know?” *Animal Behaviour* 61 (1): 139–51. doi:10.1006/anbe.2000.1518.

———. 2006. “Chimpanzees Deceive a Human Competitor by Hiding.” *Cognition* 101 (3): 495–514. doi:10.1016/j.cognition.2005.01.011.

Hare, Brian, Alicia P. Melis, Vanessa Woods, Sara Hastings, and Richard Wrangham. 2007. “Tolerance Allows Bonobos to Outperform Chimpanzees on a Cooperative Task.” *Current Biology: CB* 17 (7): 619–23. doi:10.1016/j.cub.2007.02.040.

Hare, Brian, Josep Call, Bryan Agnetta, and Michael Tomasello. 2000. “Chimpanzees Know What Conspecifics Do and Do Not See.” *Animal Behaviour* 59 (4): 771–85. doi:10.1006/anbe.1999.1377.

Harenski, Carla L., Olga Antonenko, Matthew Shane, and Kent A. Kiehl. 2009. “A Functional Imaging Investigation of Moral Deliberation and Moral Intuition.” *NeuroImage* 49 (3): 2707–16. doi:10.1016/j.neuroimage.2009.10.062.

Hatfield, Elaine, John T. Cacioppo, and Richard L. Rapson. 1993. “Emotional Contagion.” *Current Directions in Psychological Science* 2 (3): 96–99.

———. 1994. *Emotional Contagion*. Cambridge: Cambridge University Press.

Hauser, Marc. 2001. *Wild Minds: What Animals Really Think*. New Ed. London: Penguin

Books.

Hayaki, Hitoshige. 1985. “Social Play of Juvenile and Adolescent Chimpanzees in the Mahale Mountains National Park, Tanzania.” *Primates* 26 (4): 343–60. doi:10.1007/BF02382452.

“Heartbreaking Moment a Dog Tries to Pull Its Friend out of the Road.” 2014. *Mail Online*. November 24. http://www.dailymail.co.uk/news/article-2847444/The-heartbreaking-moment-young-dog-tries-pull-companion-road-left-dead-hit-car.html.

Herrera Guevara, Asunción. 2012. “Justicia Para Con Los Animales.” *Dilemata* 9: 83–87.

Heyes, Cecilia. 1998. “Theory of Mind in Nonhuman Primates.” *The Behavioral and Brain Sciences* 21 (1): 101–14; discussion 115–48.

———. 2014. “Animal Mindreading: What’s the Problem?” *Psychonomic Bulletin & Review*, August, 1–15. doi:10.3758/s13423-014-0704-4.

Hoffman, Martin L. 1990. “Empathy and Justice Motivation.” *Motivation and Emotion* 14 (2): 151–72. doi:10.1007/BF00991641.

Horner, Victoria, J. Devyn Carter, Malini Suchak, and Frans B.M. de Waal. 2011. “Spontaneous Prosocial Choice by Chimpanzees.” *Proceedings of the National Academy of Sciences* 108 (33): 13847–13851. doi:10.1073/pnas.1111088108.

Horta, Oscar. 2012. “Motivación Sin Evaluación: Definiendo La Subjetividad Moral.” *Dilemata* 0 (9): 89–104.

Hume, David. (1739) 1992. *A Treatise of Human Nature.* In *Philosophical Works*, edited by Thomas Hill Green and Thomas Hodge Grose. Vol. 2. 4 vols. London: Scientia Verlag.

Hurley, Susan, and Matthew Nudds. 2006. “The Questions of Animal Rationality: Theory and Evidence.” In *Rational Animals?*, edited by Susan Hurley and Matthew Nudds. Oxford: Oxford University Press.

Hursthouse, Rosalind. 2013. “Virtue Ethics.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2013. http://plato.stanford.edu/archives/fall2013/entries/ethics-virtue/.

Hutto, Daniel D. 2012. *Folk Psychological Narratives: The Sociocultural Basis of Understanding Reasons.* Cambridge, MA: MIT Press.

Hutto, Daniel D., and Shaun Gallagher. 2008. “Understanding Others through Primary Interaction and Narrative Practice.” In *The Shared Mind: Perspectives on Intersubjectivity*, edited by Jordan Zlatev, Timothy P. Racine, Chris Sinha, and Esa Itkonen, 12: 17–38. Converging Evidence in Language and Communication Research. Amsterdam; Philadelphia, PA: John Benjamins Publishing Company.

Ickes, William. 2009. “Empathic Accuracy: Its Links to Clinical, Cognitive, Developmental, Social, and Physiological Psychology.” In *The Social Neuroscience of Empathy*, edited by Jean Decety and William Ickes, 57–70. Cambridge, MA: MIT Press.

Jabbi, Mbemba, Marte Swart, and Christian Keysers. 2007. “Empathy for Positive and Negative Emotions in the Gustatory Cortex.” *NeuroImage* 34 (4): 1744–53. doi:10.1016/j.neuroimage.2006.10.032.

Jensen, Keith, Josep Call, and Michael Tomasello. 2007. “Chimpanzees Are Vengeful but Not Spiteful.” *Proceedings of the National Academy of Sciences* 104 (32): 13046–50. doi:10.1073/pnas.0705555104.

Jeon, Daejong, Sangwoo Kim, Mattu Chetana, Daewoong Jo, H. Earl Ruley, Shih-Yao Lin, Dania Rabah, Jean-Pierre Kinet, and Hee-Sup Shin. 2010. “Observational Fear Learning Involves Affective Pain System and Cav1.2 Ca2+ Channels in ACC.” *Nature Neuroscience* 13 (4): 482–88. doi:10.1038/nn.2504.

Joly-Mascheroni, Ramiro M., Atsushi Senju, and Alex J. Shepherd. 2008. “Dogs Catch Human Yawns.” *Biology Letters* 4 (5): 446–48. doi:10.1098/rsbl.2008.0333.

Jones, Amanda C., and Robert A. Josephs. 2006. “Interspecies Hormonal Interactions between Man and the Domestic Dog (Canis Familiaris).” *Hormones and Behavior* 50 (3): 393–400. doi:10.1016/j.yhbeh.2006.04.007.

Kagan, Jerome. 2000. “Human Morality Is Distinctive.” *Journal of Consciousness Studies* 7 (1-2): 46–48.

Kaminski, Juliane, Josep Call, and Michael Tomasello. 2008. “Chimpanzees Know What Others Know, but Not What They Believe.” *Cognition* 109 (2): 224–34. doi:10.1016/j.cognition.2008.08.010.

Kikusui, Takefumi, Shu Takigami, Yukari Takeuchi, and Yuji Mori. 2001. “Alarm Pheromone Enhances Stress-Induced Hyperthermia in Rats.” *Physiology & Behavior* 72 (1–2): 45–50. doi:10.1016/S0031-9384(00)00370-X.

Kikusui, Takefumi, James T. Winslow, and Yuji Mori. 2006. “Social Buffering: Relief from Stress and Anxiety.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 361 (1476): 2215–28. doi:10.1098/rstb.2006.1941.

Knapska, Ewelina, Evgeni Nikolaev, Pawel Boguszewski, Grazyna Walasek, Janusz Blaszczyk, Leszek Kaczmarek, and Tomasz Werka. 2006. “Between-Subject Transfer of Emotional Information Evokes Specific Pattern of Amygdala Activation.” *Proceedings of the National Academy of Sciences* 103 (10): 3858–62. doi:10.1073/pnas.0511302103.

Knobe, Joshua, and Shaun Nichols. 2008. “An Experimental Philosophy Manifesto.” In *Experimental Philosophy*, edited by Joshua Knobe and Shaun Nichols, 3–16. Oxford; New York, NY: Oxford University Press.

Korsgaard, Christine M. 2006. “Morality and the Distinctiveness of Human Action.” In *Primates and Philosophers: How Morality Evolved*, edited by Stephen Macedo and Josiah Ober, 98–119. Princeton, NJ: Princeton University Press.

Kret, Mariska E., Masaki Tomonaga, and Tetsuro Matsuzawa. 2014. “Chimpanzees and Humans Mimic Pupil-Size of Conspecifics.” *PLoS ONE* 9 (8): e104886. doi:10.1371/journal.pone.0104886.

Kuczaj, Stan, Karissa Tranel, Marie Trone, and Heather Hill. 2001. “Are Animals Capable of Deception or Empathy? Implications for Animal Consciousness and Animal Welfare.” *Animal Welfare. Special Issue* 10: 161–73.

Kunz, Thomas H., A. L. Allgaier, John Seyjagat, and Rick Caligiuri. 1994. “Allomaternal Care: Helper-Assisted Birth in the Rodrigues Fruit Bat, Pteropus Rodricensis (Chiroptera: Pteropodidae).” *Journal of Zoology* 232 (4): 691–700. doi:10.1111/j.1469-7998.1994.tb04622.x.

Kuroshima, Hika, Kazuo Fujita, Ikuma Adachi, Kana Iwata, and Akira Fuyuki. 2003. “A Capuchin Monkey (Cebus Apella) Recognizes When People Do and Do Not Know the Location of Food.” *Animal Cognition* 6 (4): 283–91. doi:10.1007/s10071-003-0184-9.

Kuroshima, Hika, Kazuo Fujita, Akira Fuyuki, and Tsuyuka Masuda. 2002. “Understanding of the Relationship between Seeing and Knowing by Tufted Capuchin Monkeys (Cebus Apella).” *Animal Cognition* 5 (1): 41–48.

Kutsukake, Nobuyuki, and Duncan L. Castles. 2004. “Reconciliation and Post-Conflict Third-Party Affiliation among Wild Chimpanzees in the Mahale Mountains, Tanzania.” *Primates; Journal of Primatology* 45 (3): 157–65. doi:10.1007/s10329-004-0082-z.

Lakshminarayanan, Venkat R., and Laurie R. Santos. 2008. “Capuchin Monkeys Are Sensitive to Others’ Welfare.” *Current Biology* 18 (21): R999–1000. doi:10.1016/j.cub.2008.08.057.

Lane, Jonathan D., Henry M. Wellman, Sheryl L. Olson, Jennifer LaBounty, and David C. R. Kerr. 2010. “Theory of Mind and Emotion Understanding Predict Moral Development in Early Childhood.” *The British Journal of Developmental Psychology* 28 (Pt 4): 871–89.

Langford, Dale J., Sara E. Crager, Zarrar Shehzad, Shad B. Smith, Susana G. Sotocinal, Jeremy S. Levenstadt, Mona Lisa Chanda, Daniel J. Levitin, and Jeffrey S. Mogil. 2006. “Social Modulation of Pain as Evidence for Empathy in Mice.” *Science* 312 (5782): 1967–70. doi:10.1126/science.1128322.

Lavery, J.J., and P.J. Foley. 1963. “Altruism or Arousal in the Rat?” *Science* 140 (3563): 172–73. doi:10.1126/science.140.3563.172.

Lenman, James. 2011. “Reasons for Action: Justification vs. Explanation.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2011. http://plato.stanford.edu/archives/win2011/entries/reasons-just-vs-expl/.

Leudar, Ivan, and Alan Costall (Eds.). 2011. *Against Theory of Mind*. Reprint. London: Palgrave Macmillan.

Lingle, Susan, and Tobias Riede. 2014. “Deer Mothers Are Sensitive to Infant Distress Vocalizations of Diverse Mammalian Species.” *The American Naturalist* 184 (4): 510–22. doi:10.1086/677677.

Lipps, Theodor. 1903. “Einfühlung, Innere Nachahmung Und Organempfindung.” *Archiv Für Gesamte Psychologie* 1: 465–519.

“Living Links | Video.” 2016. Accessed March 14. http://www.emory.edu/LIVING_LINKS/media/video.shtml.

Longueira Monelos, Angel. 2012. “Crítica a La Pregunta ‘¿Pueden Los Animales Ser Morales?’” *Dilemata* 9: 65–68.

Lucke, Joseph F, and C. Daniel Batson. 1980. “Response Suppression to a Distressed Conspecific: Are Laboratory Rats Altruistic?” *Journal of Experimental Social Psychology* 16 (3): 214–27. doi:10.1016/0022-1031(80)90065-7.

Lurz, Robert W. 2009. “If Chimpanzees Are Mindreaders, Could Behavioral Science Tell? Toward a Solution of the Logical Problem.” *Philosophical Psychology* 22 (3): 305–28. doi:10.1080/09515080902970673.

———. 2011. *Mindreading Animals: The Debate over What Animals Know about Other Minds.* Cambridge, MA: MIT Press.

Machan, Tibor R. 2002. “Why Human Beings May Use Animals.” *The Journal of Value Inquiry* 36 (1): 9–16. doi:10.1023/A:1014993828953.

Mancini, Giada, Pier Francesco Ferrari, and Elisabetta Palagi. 2013. “Rapid Facial Mimicry in Geladas.” *Scientific Reports* 3 (1527). doi:10.1038/srep01527.

Massen, Jorg J. M., Lisette M. Van Den Berg, Berry M. Spruijt, and Elisabeth H. M. Sterck. 2012. “Inequity Aversion in Relation to Effort and Relationship Quality in Long-Tailed Macaques (Macaca Fascicularis).” *American Journal of Primatology* 74 (2): 145–56. doi:10.1002/ajp.21014.

Masserman, Jules, Stanley Wechkin, and William Terris. 1964. “‘Altruistic’ Behaviour in Rhesus Monkeys.” *American Journal of Psychiatry* 121 (6): 584–85.

McCloskey, Henry J. 1979. “Moral Rights and Animals.” *Inquiry* 22 (1-4): 23–54. doi:10.1080/00201747908601865.

———. 1987. “The Moral Case for Experimentation on Animals.” *The Monist* 70 (1): 64–82.

McComb, Karen, Lucy Baker, and Cynthia Moss. 2006. “African Elephants Show High Levels of Interest in the Skulls and Ivory of Their Own Species.” *Biology Letters* 2 (1): 26–28. doi:10.1098/rsbl.2005.0400.

Melis, Alicia P., Josep Call, and Michael Tomasello. 2006. “Chimpanzees (Pan Troglodytes) Conceal Visual and Auditory Information from Others.” *Journal of Comparative Psychology* 120 (2): 154–62. doi:10.1037/0735-7036.120.2.154.

Melis, Alicia P., Brian Hare, and Michael Tomasello. 2006. “Engineering Cooperation in Chimpanzees: Tolerance Constraints on Cooperation.” *Animal Behaviour* 72: 275–86.

Meltzoff, Andrew N., and M. Keith Moore. 1977. “Imitation of Facial and Manual Gestures by Human Neonates.” *Science* 198 (4312): 74–78.

Merola, Isabella, Emanuela Prato-Previde, M. Lazzaroni, and Sarah Marshall-Pescini. 2013. “Dogs’ Comprehension of Referential Emotional Expressions: Familiar People and Familiar Emotions Are Easier.” *Animal Cognition* 17 (2): 373–85. doi:10.1007/s10071-013-0668-1.

Messenger, Stephen. 2014. “Why This Video Of A Beluga Whale ‘Playing’ With Children Is Actually Very Sad.” *The Dodo*. August 22. https://www.thedodo.com/why-this-video-of-a-beluga-wha-685343078.html.

Michael, John. 2014. “Towards a Consensus about the Role of Empathy in Interpersonal Understanding.” *Topoi* 33 (1): 157–72. doi:10.1007/s11245-013-9204-9.

Micheletta, Jérôme, Jamie Whitehouse, Lisa A. Parr, and Bridget M. Waller. 2015. “Facial Expression Recognition in Crested Macaques (Macaca Nigra).” *Animal Cognition* 18 (4): 985–90. doi:10.1007/s10071-015-0867-z.

Michelle Cotton. 2014. Accessed March 14 2016. *Whale CT...Belugas Tease Our Three Kids*. https://www.youtube.com/watch?v=Hh84Oe8JxUQ.

Miller, Robert E., John V. Murphy, and I. Arthur Mirsky. 1959. “Relevance of Facial Expression and Posture as Cues in Communication of Affect between Monkeys.” *A.M.A. Archives of General Psychiatry* 1 (November): 480–88.

Miller, Robert E., James H. Banks Jr., and Nobuya Ogawa. 1963. “Role of Facial Expression in ‘Cooperative-Avoidance Conditioning’ in Monkeys.” *The Journal of Abnormal and Social Psychology* 67 (1): 24–30. doi:10.1037/h0044018.

Miller, Robert E., William F. Caul, and I. Arthur Mirsky. 1967. “Communication of Affects between Feral and Socially Isolated Monkeys.” *Journal of Personality and Social Psychology* 7 (3): 231–39.

Mitchell, Robert W., Nicholas S. Thompson, and H. Lyn Miles (Eds.). 1997. *Anthropomorphism, Anecdotes, and Animals*. Albany, NY: SUNY Press.

"Monkey Saves Dying Friend at Indian Train Station." 2014. *The Guardian*. December 22. http://www.theguardian.com/world/video/2014/dec/22/monkey-saves-dying-friend-train-station-india-video.

Morimoto, Yo, and Kazuo Fujita. 2012. "Capuchin Monkeys (Cebus Apella) Use Conspecifics' Emotional Expressions to Evaluate Emotional Valence of Objects." *Animal Cognition* 15 (3): 341–47. doi:10.1007/s10071-011-0458-6.

Müller, Corsin A., Kira Schmitt, Anjuli L. A. Barber, and Ludwig Huber. 2015. "Dogs Can Discriminate Emotional Expressions of Human Faces." *Current Biology* 25 (5): 601–5. doi:10.1016/j.cub.2014.12.055.

Nagasawa, Miho, Kensuke Murai, Kazutaka Mogi, and Takefumi Kikusui. 2011. "Dogs Can Discriminate Human Smiling Faces from Blank Expressions." *Animal Cognition* 14 (4): 525–33. doi:10.1007/s10071-011-0386-5.

Nakashima, Satoshi F., Masatoshi Ukezono, Hiroshi Nishida, Ryunosuke Sudo, and Yuji Takano. 2015. "Receiving of Emotional Signal of Pain from Conspecifics in Laboratory Rats." *Royal Society Open Science* 2 (4): 140381. doi:10.1098/rsos.140381.

Newen, Albert, and Andreas Bartels. 2007. "Animal Minds and the Possession of Concepts." *Philosophical Psychology* 20 (3): 283–308. doi:10.1080/09515080701358096.

Nichols, Shaun. 2001. "Mindreading and the Cognitive Architecture Underlying Altruistic Motivation." *Mind & Language* 16 (4): 425–55. doi:10.1111/1468-0017.00178.

Nichols, Shaun, and Stephen P. Stich. 2003. *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. Oxford: Clarendon Press.

Nowbahari, Elise, Alexandra Scohier, Jean-Luc Durand, and Karen L. Hollis. 2009. "Ants, Cataglyphis Cursor, Use Precisely Directed Rescue Behavior to Free Entrapped Relatives." *PLoS ONE* 4 (8): e6573. doi:10.1371/journal.pone.0006573.

Nussbaum, Martha C. 2004. "Beyond 'Compassion and Humanity': Justice for Nonhuman Animals." In *Animal Rights: Current Debates and New Directions*, edited by Cass R. Sunstein and Martha C. Nussbaum, 299–320. New York, NY: Oxford University Press.

———. 2007. *Frontiers of Justice: Disability, Nationality, Species Membership*. New Ed. Cambridge, MA: Harvard University Press.

Oberman, Lindsay M., and Vilayanur S. Ramachandran. 2007. "The Simulating Social Mind: The Role of the Mirror Neuron System and Simulation in the Social and Communicative Deficits of Autism Spectrum Disorders." *Psychological Bulletin* 133 (2): 310–27. doi:10.1037/0033-2909.133.2.310.

O'Connell, Sanjida M. 1995. "Empathy in Chimpanzees: Evidence for Theory of Mind?" *Primates* 36 (3): 397–410. doi:10.1007/BF02382862.

Packer, Craig. 1977. “Reciprocal Altruism in Papio Anubis.” *Nature* 265 (5593): 441–43. doi:10.1038/265441a0.

Palagi, Elisabetta, Alessia Leone, Giada Mancini, and Pier Francesco Ferrari. 2009. “Contagious Yawning in Gelada Baboons as a Possible Expression of Empathy.” *Proceedings of the National Academy of Sciences* 106 (46): 19262–67. doi:10.1073/pnas.0910891106.

Palagi, Elisabetta, and Giada Cordoni. 2009. “Postconflict Third-Party Affiliation in Canis Lupus: Do Wolves Share Similarities with the Great Apes?” *Animal Behaviour* 78 (4): 979–86. doi:10.1016/j.anbehav.2009.07.017.

Palagi, Elisabetta, Stefania Dall’Olio, Elisa Demuru, and Roscoe Stanyon. 2014. “Exploring the Evolutionary Foundations of Empathy: Consolation in Monkeys.” *Evolution and Human Behavior* 35 (4): 341–49. doi:10.1016/j.evolhumbehav.2014.04.002.

Palagi, Elisabetta, Velia Nicotra, and Giada Cordoni. 2015. “Rapid Mimicry and Emotional Contagion in Domestic Dogs.” *Royal Society Open Science* 2 (12): 150505. doi:10.1098/rsos.150505.

Palagi, Elisabetta, and Ivan Norscia. 2013. “Bonobos Protect and Console Friends and Kin.” *PLoS ONE* 8 (11): e79290. doi:10.1371/journal.pone.0079290.

Palagi, Elisabetta, Ivan Norscia, and Elisa Demuru. 2014. “Yawn Contagion in Humans and Bonobos: Emotional Affinity Matters More than Species.” *PeerJ* 2 (August): e519. doi:10.7717/peerj.519.

Palagi, Elisabetta, Tommaso Paoli, and Silvana Borgognini Tarli. 2004. “Reconciliation and Consolation in Captive Bonobos (Pan Paniscus).” *American Journal of Primatology* 62 (1): 15–30. doi:10.1002/ajp.20000.

Panksepp, Jaak. 2004. *Affective Neuroscience: The Foundations of Human and Animal Emotions*. 1st ed. New York, NY: Oxford University Press.

Park, Kyum J., Hawsun Sohn, Yong R. An, Dae Y. Moon, Seok G. Choi, and Doo H. An. 2012. “An Unusual Case of Care-Giving Behavior in Wild Long-Beaked Common Dolphins (Delphinus Capensis) in the East Sea.” *Marine Mammal Science* 29 (4): E508–14. doi:10.1111/mms.12012.

Parr, Lisa A. 2001. “Cognitive and Physiological Markers of Emotional Awareness in Chimpanzees (Pan Troglodytes).” *Animal Cognition* 4 (3-4): 223–29. doi:10.1007/s100710100085.

———. 2003. “The Discrimination of Faces and Their Emotional Content by Chimpanzees (Pan Troglodytes).” *Annals of the New York Academy of Sciences* 1000 (December): 56–78.

Parr, Lisa A., and Matthew Heintz. 2009. "Facial Expression Recognition in Rhesus Monkeys, Macaca Mulatta." *Animal Behaviour* 77 (6): 1507–13. doi:10.1016/j.anbehav.2009.02.024.

Parr, Lisa A., William D. Hopkins, and Frans B.M. de Waal. 1998. "The Perception of Facial Expressions By Chimpanzees, *Pan Troglodytes*." *Evolution of Communication* 2 (1): 1–23. doi:10.1075/eoc.2.1.02par.

Parr, Lisa A., James T. Winslow, William D. Hopkins, and Frans B.M. de Waal. 2000. "Recognizing Facial Cues: Individual Discrimination by Chimpanzees (Pan Troglodytes) and Rhesus Monkeys (Macaca Mulatta)." *Journal of Comparative Psychology (Washington, D.C. : 1983)* 114 (1): 47–60.

Paukner, Annika, and James R. Anderson. 2006. "Video-Induced Yawning in Stumptail Macaques (Macaca Arctoides)." *Biology Letters* 2 (1): 36–38. doi:10.1098/rsbl.2005.0411.

Penn, Derek C., Keith J. Holyoak, and Daniel J. Povinelli. 2008. "Darwin's Mistake: Explaining the Discontinuity between Human and Nonhuman Minds." *The Behavioral and Brain Sciences* 31 (2): 109–30; discussion 130–78. doi:10.1017/S0140525X08003543.

Penn, Derek C., and Daniel J. Povinelli. 2007. "On the Lack of Evidence That Non-Human Animals Possess Anything Remotely Resembling a 'Theory of Mind.'" *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 362 (1480): 731–44. doi:10.1098/rstb.2006.2023.

———. 2013. "The Comparative Delusion: The 'behavioristic'/ 'mentalistic' Dichotomy in Comparative Theory of Mind Research." In *Agency and Joint Attention*, edited by Herbert A. Terrace and Janet Metcalfe, 62–81. New York, NY: Oxford University Press.

Peterson, Gregory R. 2000. "God, Genes, and Cognizing Agents." *Zygon* 35 (3): 469–80.

Pfeifer, Jennifer H., and Mirella Dapretto. 2009. "'Mirror, Mirror, in My Mind': Empathy, Interpersonal Competence, and the Mirror Neuron System." In *The Social Neuroscience of Empathy*, edited by Jean Decety and William Ickes, 183–98. Social Neuroscience Series. Boston, MA: MIT Press.

Piazzarollo Loureiro, Carolina, and Debora de Hollanda Souza. 2013. "The Relationship between Theory of Mind and Moral Development in Preschool Children." *Paidéia (Ribeirão Preto)* 23 (54): 93–101. doi:10.1590/1982-43272354201311.

"Planning of the Apes: Zoo Chimp Plots Rock Attacks on Visitors." 2014. Accessed December 12. http://www.scientificamerican.com/article/chimpanzee-plans-throws-stones-zoo/.

Plotnik, Joshua M., and Frans B.M. de Waal. 2014. "Asian Elephants (Elephas Maximus)

Reassure Others in Distress." *PeerJ* 2 (e278). doi:10.7717/peerj.278.

Plotnik, Joshua M., Richard Lair, Wirot Suphachoksahakun, and Frans B.M. de Waal. 2011. "Elephants Know When They Need a Helping Trunk in a Cooperative Task." *Proceedings of the National Academy of Sciences*, March, 201101765. doi:10.1073/pnas.1101765108.

Pluhar, Evelyn B. 1995. *Beyond Prejudice: The Moral Significance of Human and Nonhuman Animals*. Durham, NC: Duke University Press.

Porter, James. 1977. "Pseudorca Stranding." *Oceans* 10: 8–15.

Povinelli, Daniel J. 1998. "Can Animals Empathize? Maybe Not." *Scientific American* 9: 67–75.

Povinelli, Daniel J., and Timothy J. Eddy. 1996. "What Young Chimpanzees Know about Seeing." *Monographs of the Society for Research in Child Development* 61 (3): i – vi, 1–152; discussion 153–91.

Povinelli, Daniel J., Kurt E. Nelson, and Sarah T. Boysen. 1990. "Inferences about Guessing and Knowing by Chimpanzees (Pan Troglodytes)." *Journal of Comparative Psychology* 104 (3): 203–10.

Povinelli, Daniel J., Helen K. Perilloux, James E. Reaux, and Donna T. Bierschwale. 1998. "Young and Juvenile Chimpanzees' (Pan Troglodytes) Reactions to Intentional versus Accidental and Inadvertent Actions." *Behavioural Processes* 42 (2-3): 205–18.

Povinelli, Daniel J, and Jennifer Vonk. 2003. "Chimpanzee Minds: Suspiciously Human?" *Trends in Cognitive Sciences* 7 (4): 157–60.

———. 2004. "We Don't Need a Microscope to Explore the Chimpanzee's Mind." *Mind and Language* 19 (1): 1–28.

Prather, Jonathan F., Susan Peters, Stephen Nowicki, and Richard Mooney. 2008. "Precise Auditory–vocal Mirroring in Neurons for Learned Vocal Communication." *Nature* 451 (7176): 305–10. doi:10.1038/nature06492.

Premack, David. 1988. "'Does the Chimpanzee Have a Theory of Mind' Revisited." In *Machiavellian Intelligence*, edited by Richard Byrne and Andrew Whiten, 160–79. New York, NY: Oxford University Press.

Premack, David, and Guy Woodruff. 1978. "Does the Chimpanzee Have a Theory of Mind?" *Behavioral and Brain Sciences* 1 (04): 515–26. doi:10.1017/S0140525X00076512.

Preston, Stephanie D., and Frans B.M. de Waal. 2002. "Empathy: Its Ultimate and Proximate Bases." *Behavioral and Brain Sciences* 25 (1): 1–20; discussion 20–71.

Range, Friederike, Lisa Horn, Zsófia Viranyi, and Ludwig Huber. 2009. "The Absence of Reward Induces Inequity Aversion in Dogs." *Proceedings of the National Academy of*

Sciences 106 (1): 340–45. doi:10.1073/pnas.0810957105.

Ravenscroft, Ian. 2010. “Folk Psychology as a Theory.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2010. http://plato.stanford.edu/archives/fall2010/entries/folkpsych-theory/.

Regan, Tom. (1983) 2004. *The Case for Animal Rights*. First Edition, Updated with a New Preface edition. Berkeley, CA: University of California Press.

Reimert, Inonge, J. Elizabeth Bolhuis, Bas Kemp, and T. Bas Rodenburg. 2013. “Indicators of Positive and Negative Emotions and Emotional Contagion in Pigs.” *Physiology & Behavior* 109 (January): 42–50. doi:10.1016/j.physbeh.2012.11.002.

———. 2015. “Emotions on the Loose: Emotional Contagion and the Role of Oxytocin in Pigs.” *Animal Cognition* 18 (2): 517–32. doi:10.1007/s10071-014-0820-6.

Rice, George E., and Priscilla Gainer. 1962. “‘Altruism’ in the Albino Rat.” *Journal of Comparative and Physiological Psychology* 55 (February): 123–25.

Rizzolatti, Giacomo, Luciano Fadiga, Vittorio Gallese, and Leonardo Fogassi. 1996. “Premotor Cortex and the Recognition of Motor Actions.” *Brain Research. Cognitive Brain Research* 3 (2): 131–41.

Rizzolatti, Giacomo, and Laila Craighero. 2004. “The Mirror-Neuron System.” *Annual Review of Neuroscience* 27: 169–92. doi:10.1146/annurev.neuro.27.070203.144230.

Romero, Teresa, Marie Ito, Atsuko Saito, and Toshikazu Hasegawa. 2014. “Social Modulation of Contagious Yawning in Wolves.” *PLoS ONE* 9 (8): e105963. doi:10.1371/journal.pone.0105963.

Romero, Teresa, Akitsugu Konno, and Toshikazu Hasegawa. 2013. “Familiarity Bias and Physiological Responses in Contagious Yawning by Dogs Support Link to Empathy.” *PLoS ONE* 8 (8): e71365. doi:10.1371/journal.pone.0071365.

Rowlands, Mark. 2009. *Animal Rights: Moral Theory and Practice*. 2 edition. Houndmills, Basingstoke; New York, NY: Palgrave Macmillan.

———. 2011. “Animals That Act for Moral Reasons.” In *Oxford Handbook of Animal Ethics*, edited by Tom L. Beauchamp and Raymond G. Frey. New York, NY: Oxford University Press.

———. 2012a. “¿Pueden los animales ser morales?” *Dilemata* 0 (9): 1–32.

———. 2012b. *Can Animals Be Moral?* New York, NY: Oxford University Press.

———. 2013. “Animals and Moral Motivation: A Response to Clement.” *Journal of Animal Ethics* 3 (1): 15–24.

———. Forthcoming 2017. “Moral Subjects.” In *The Routledge Handbook of Philosophy of Animal Minds*, edited by Kristin Andrews and Jacob Beck. London: Routledge.

Rowlands, Mark, and Susana Monsó. Forthcoming 2016. "Animals as Reflexive Thinkers: The Aponoian Paradigm." In *The Oxford Handbook of Animal Studies*, edited by Linda Kalof. Oxford University Press. doi:10.1093/oxfordhb/9780199927142.013.15.

Rutte, Claudia, and Michael Taborsky. 2007. "Generalized Reciprocity in Rats." *PLoS Biology* 5 (7): 1421–25. doi:10.1371/journal.pbio.0050196.

Sagi, Abraham, and Martin L. Hoffman. 1976. "Empathic Distress in the Newborn." *Developmental Psychology* 12 (2): 175–76. doi:10.1037/0012-1649.12.2.175.

Sánchez Suárez, Walter. 2012. "¿Más Que Pacientes Morales?" *Dilemata* 0 (9): 131–34.

Santos, Laurie R., Aaron G. Nissen, and Jonathan A. Ferrugia. 2006. "Rhesus Monkeys, Macaca Mulatta, Know What Others Can and Cannot Hear." *Animal Behaviour* 71 (5): 1175–81. doi:10.1016/j.anbehav.2005.10.007.

Sapontzis, Steve F. 1987. *Morals, Reason, and Animals*. Philadelphia, PA: Temple University Press.

Sato, Nobuya, Ling Tan, Kazushi Tate, and Maya Okada. 2015. "Rats Demonstrate Helping Behavior toward a Soaked Conspecific." *Animal Cognition* 18 (5): 1039–47. doi:10.1007/s10071-015-0872-2.

Savage-Rumbaugh, E. Sue, Duane M. Rumbaugh, and Sally Boysen. 1978. "Sarah's Problems of Comprehension." *Behavioral and Brain Sciences* 1 (04): 555–57. doi:10.1017/S0140525X0007655X.

Searle, John R. 1994. "Animal Minds." *Midwest Studies In Philosophy* 19 (1): 206–19. doi:10.1111/j.1475-4975.1994.tb00286.x.

Seed, Amanda M., Nicola S. Clayton, and Nathan J. Emery. 2007. "Postconflict Third-Party Affiliation in Rooks, Corvus Frugilegus." *Current Biology: CB* 17 (2): 152–58. doi:10.1016/j.cub.2006.11.025.

Seyfarth, Robert M., and Dorothy L. Cheney. 1984. "Grooming, Alliances and Reciprocal Altruism in Vervet Monkeys." *Nature* 308 (5959): 541–43.

Silk, Joan B., Sarah F. Brosnan, Jennifer Vonk, Joseph Henrich, Daniel J. Povinelli, Amanda S. Richardson, Susan P. Lambeth, Jenny Mascaro, and Steven J. Schapiro. 2005. "Chimpanzees Are Indifferent to the Welfare of Unrelated Group Members." *Nature* 437 (7063): 1357–59. doi:10.1038/nature04243.

Singer, Tania. 2006. "The Neuronal Basis and Ontogeny of Empathy and Mind Reading: Review of Literature and Implications for Future Research." *Neuroscience & Biobehavioral Reviews* 30 (6): 855–63. doi:10.1016/j.neubiorev.2006.06.011.

Slocombe, Katie E., Simon W. Townsend, and Klaus Zuberbühler. 2009. "Wild Chimpanzees (Pan Troglodytes Schweinfurthii) Distinguish between Different Scream Types:

Evidence from a Playback Study." *Animal Cognition* 12 (3): 441–49. doi:10.1007/s10071-008-0204-x.

Smith, Adam. (1759) 1982. *The Theory of Moral Sentiments*. Edited by David D. Raphael and Alexander L. Macfie. Indianapolis, IN: Liberty Fund.

Smith, Adam S., and Zuoxin Wang. 2014. "Hypothalamic Oxytocin Mediates Social Buffering of the Stress Response." *Biological Psychiatry* 76 (4): 281–88. doi:10.1016/j.biopsych.2013.09.017.

Smith, Amy Victoria, Leanne Proops, Kate Grounds, Jennifer Wathan, and Karen McComb. 2016. "Functionally Relevant Responses to Human Facial Expressions of Emotion in the Domestic Horse (Equus Caballus)." *Biology Letters* 12 (2): 20150907. doi:10.1098/rsbl.2015.0907.

Smith, J. David, Michael J. Beran, Justin J. Couchman, Mariana V. C. Coutinho, and Joseph B. Boomer. 2009. "Animal Metacognition: Problems and Prospects." *Comparative Cognition & Behavior Reviews* 4: 40–53. doi:10.3819/ccbr.2009.40004.

Spaulding, Shannon. 2015. "On Direct Social Perception." *Consciousness and Cognition* 36 (November): 472–82. doi:10.1016/j.concog.2015.01.003.

Stander, P. E. 1992. "Cooperative Hunting in Lions: The Role of the Individual." *Behavioral Ecology and Sociobiology* 29 (6): 445–54. doi:10.1007/BF00170175.

Stich, Stephen P. 1979. "Do Animals Have Beliefs?" *Australasian Journal of Philosophy* 57 (March): 15–28.

Stueber, Karsten. 2014. "Empathy." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2014. http://plato.stanford.edu/archives/win2014/entries/empathy/.

Sümegi, Zsófia, Katalin Oláh, and József Topál. 2014. "Emotional Contagion in Dogs as Measured by Change in Cognitive Task Performance." *Applied Animal Behaviour Science* 160 (November): 106–15. doi:10.1016/j.applanim.2014.09.001.

Tafalla, Marta. 2012. "¿Son algunos mamíferos sujetos proto-morales? Tres observaciones y una paradoja." *Dilemata* 0 (9): 53–63.

Tamames, Kepa. 2012. "¿Les Importa a Los Animales Ser (o No Ser) Sujetos Morales?" *Dilemata* 0 (9): 141–51.

Titchener, Edward B. 1909. *Lectures on the Experimental Psychology of the Thought-Processes*. New York, NY: The Macmillan company.

Tomasello, Michael, Brian Hare, and Bryan Agnetta. 1999. "Chimpanzees, Pan Troglodytes, Follow Gaze Direction Geometrically." *Animal Behaviour* 58: 769–77.

Tomasello, Michael, Brian Hare, and Tara Fogleman. 2001. "The Ontogeny of Gaze Following in Chimpanzees, Pan Troglodytes, and Rhesus Macaques, Macaca Mulatta." *Animal Behaviour* 61 (2): 335–43. doi:10.1006/anbe.2000.1598.

Torres Aldave, Mikel. 2012. "¿Es 'Sujeto Moral' Un Concepto Inútil (y Peligroso)? Comentario a «¿Pueden Los Animales Ser Morales?»." *Dilemata* 0 (9): 105–21.

Trivers, Robert L. 1971. "The Evolution of Reciprocal Altruism." *Quarterly Review of Biology* 46: 35–57.

Vasconcelos, Marco, Karen Hollis, Elise Nowbahari, and Alex Kacelnik. 2012. "Pro-Sociality without Empathy." *Biology Letters* 8 (6): 910–12. doi:10.1098/rsbl.2012.0554.

Villacorta, Joana Charterina. 2012. "¿Puede Nuestra Consciencia Moral Asumir La Moralidad Animal?" *Dilemata* 0 (9): 41–52.

Waller, Bruce N. 1997. "What Rationality Adds to Animal Morality." *Biology and Philosophy* 12 (3): 341–56.

Warneken, Felix, Brian Hare, Alicia P. Melis, Daniel Hanus, and Michael Tomasello. 2007. "Spontaneous Altruism by Chimpanzees and Young Children." *PLoS Biology* 5 (7): e184. doi:10.1371/journal.pbio.0050184.

Warneken, Felix, and Michael Tomasello. 2006. "Altruistic Helping in Human Infants and Young Chimpanzees." *Science* 311 (5765): 1301–3. doi:10.1126/science.1121448.

Wascher, Claudia A. F., Isabella B. R. Scheiber, and Kurt Kotrschal. 2008. "Heart Rate Modulation in Bystanding Geese Watching Social and Non-Social Events." *Proceedings of the Royal Society of London B: Biological Sciences* 275 (1643): 1653–59. doi:10.1098/rspb.2008.0146.

Watanabe, Shigeru, and Kunihiko Ono. 1986. "An Experimental Analysis of 'Empathic' Response: Effects of Pain Reactions of Pigeon upon Other Pigeon's Operant Behavior." *Behavioural Processes* 13 (3): 269–77. doi:10.1016/0376-6357(86)90089-6.

Watson, Duncan M., and David B. Croft. 1996. "Age-Related Differences in Playfighting Strategies of Captive Male Red-Necked Wallabies (Macropus Rufogriseus Banksianus)." *Ethology* 102 (2): 336–46. doi:10.1111/j.1439-0310.1996.tb01129.x.

Wechkin, Stanley, Jules Masserman, and William Terris. 1964. "Shock to a Conspecific as an Aversive Stimulus." *Psychonomic Science* 1: 17–18.

Wilkinson, Gerald S. 1984. "Reciprocal Food Sharing in the Vampire Bat." *Nature* 308 (5955): 181–84. doi:10.1038/308181a0.

Wilson, Edward O., Nathaniel I. Durlach, and Louise M. Roth. 1958. "Chemical Releaser of Necrophoric Behavior in Ants." *Psyche: A Journal of Entomology* 65 (4): 108–14. doi:10.1155/1958/69391.

Wimmer, Heinz, and Josef Perner. 1983. "Beliefs about Beliefs: Representation and

Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception." *Cognition* 13: 103–28.

Wispé, Lauren. 1986. "The Distinction between Sympathy and Empathy: To Call Forth a Concept, a Word Is Needed." *Journal of Personality and Social Psychology* 50 (2): 314–21. doi:10.1037/0022-3514.50.2.314.

Woodruff, Guy, and David Premack. 1979. "Intentional Communication in the Chimpanzee: The Development of Deception." *Cognition* 7 (4): 333–62.

Woolfolk, Robert L., John M. Doris, and John M. Darley. 2006. "Identification, Situational Constraint, and Social Cognition: Studies in the Attribution of Moral Responsibility." *Cognition* 100 (2): 283–301.

Yong, Min Hooi, and Ted Ruffman. 2014. "Emotional Contagion: Dogs and Humans Show a Similar Physiological Response to Human Infant Crying." *Behavioural Processes* 108 (October): 155–65. doi:10.1016/j.beproc.2014.10.006.

Young, Liane, Fiery Cushman, Marc Hauser, and Rebecca Saxe. 2007. "The Neural Basis of the Interaction between Theory of Mind and Moral Judgment." *Proceedings of the National Academy of Sciences* 104 (20): 8235–40. doi:10.1073/pnas.0701408104.

Young, Liane, and Rebecca Saxe. 2008. "The Neural Basis of Belief Encoding and Integration in Moral Judgment." *NeuroImage* 40 (4): 1912–20. doi:doi:10.1016/j. neuroimage.2008.01.057.

Young, Liane, and Adam Waytz. 2013. "Mind Attribution Is for Morality." In *Understanding Other Minds: Perspectives from Developmental Social Neuroscience*, edited by Simon Baron-Cohen, Michael Lombardo, and Helen Tager-Flusberg, 93–103. New York, NY: Oxford University Press.

Zahavi, Dan. 2014. "Empathy and Other-Directed Intentionality." *Topoi* 33: 129–42. doi:10.1007/s11245-013-9197-4.

Zahavi, Dan, and Søren Overgaard. 2012. "Empathy without Isomorphism: A Phenomenological Account." In *Empathy: From Bench to Bedside*, edited by Jean Decety, 3–20. Cambridge, MA: MIT Press.