5-29-2015

Empathy's Significance for the Moral Status of Nonhuman Animals

Sarah Katherine Vincent

EMPATHY’S SIGNIFICANCE FOR THE MORAL STATUS OF NONHUMAN ANIMALS

by

Sarah K. Vincent

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Major: Philosophy

The University of Memphis

August 2015

Abstract

Vincent, Sarah Katherine. PhD. The University of Memphis. August, 2015. Empathy’s Significance for the Moral Status of Nonhuman Animals. Major Professors: Dr. Stephan Blatti and Dr. Shaun Gallagher.

The familiar divide between moral agents and patients in philosophical ethics seems unable to accommodate a growing body of empirical evidence that indicates the presence of moral behavior in nonhuman animals. This dissertation develops the idea that nonhuman animals might be moral subjects without necessarily being agents. Broadly, moral subjects are beings who can act for moral reasons, while moral agents can additionally scrutinize their motivations to act. A three-way distinction between moral patients, moral subjects, and moral subjects who are also agents provides the conceptual space needed to analyze and explain the pro-social, altruistic, and other-regarding behaviors evident in members of other species.

Specifically, in what follows, I argue that we ought to think of some nonhuman animals as beings who can act morally through the exercise of their capacities for *empathy* (explicated in terms of an interaction theory model) and *adaptivity* (in response to feedback about their behaviors from conspecifics or interspecies members of their community). I defend the claim that those nonhuman animals whose behavior indicates the exercise of these capacities are moral subjects. Approaching matters in this way allows us to avoid the problems associated with understanding nonhuman animals as moral agents, while also acknowledging (in keeping with recent empirical findings) that there may be other ways of being moral that are not reducible to being an object of moral concern.

Table of Contents

1

Introduction

As human animals, we are surrounded by other animal species with which we share our global world, our specific environments, and at times our individual lives. It is in our meeting of individuals from other species that we are forced to confront two crucial questions – What makes us different from them? What makes us similar to them?

Depending on whom you consult, some will argue that there appears to be a sharp divide between humans and the rest of the animal kingdom. Perhaps it is a matter of linguistic capacities – with respect to either the profound uniqueness of human language or its relative complexity when juxtaposed against the communicative capacities of other animals. There is the well-known remark by Ludwig Wittgenstein: "If a lion could talk, we wouldn't be able to understand it."[1]

Or perhaps, more strongly, other animals are machines "hav[ing] no intelligence at all," guided only by "nature which acts in them."[2] These words are of course those of René Descartes, one of several modern philosophers who saw other animals as profoundly other, as unlike us in many important respects. For one reason or another, the idea that a strict divide between humans and all other animals exists has been widely granted – and has often been accompanied by a commitment to the superiority of humans. The idea that humans are unique among other animals is then hardly new, and in

[1] Ludwig Wittgenstein, *Philosophical Investigations*, fourth edition (West Sussex: Wiley-Blackwell, 2009), 235.

[2] René Descartes, *Selected Philosophical Writings*, 1988 reprint, trans. John Cottingham, Robert Stoothoff, and Dugald Murdoch (Cambridge: Cambridge University Press, 1999), 45.

fact, has been the dominant view among philosophers and scholars from diverse disciplines.

Alternatively, some thinkers will contend that other animals should be regarded as kindred kinds of beings – beings who simply inhabit different kinds of bodies but who have at least some of the capacities historically thought to belong only to humans. Consider Tom Regan's argument that all animals (not just humans) are "subjects-of-a-life" – beings who have beliefs, desires, perception, memory, affectivity, sentience, and "an individual welfare" in virtue of the fact that their lives "far[e] well or ill for them."[3]

Moreover, perhaps it is our similarities that create the possibility for interspecies relationships and substantive communication. Consider Donna Haraway's work on dogs. Haraway adamantly defends the idea that certain similarities exist that can overcome the differences between humans and dogs, making possible meaningful relationships between members of different species. While relationships between individuals from companion species can be "messy,"[4] she contends that we can and should learn "how to see who the dogs are and hear what they are telling us, not in bloodless abstraction, but in one-on-one relationship, in otherness-in-connection."[5] These are just two examples of a slowly growing body of contemporary philosophers and researchers who are beginning to understand other animals as far more like us – or rather to see humans as far more like them – than has been traditionally believed.

[3] Tom Regan, *The Case for Animal Rights*, 1983 reprint (Berkeley: University of California Press, 2004), 243.

[4] Donna Haraway, *The Companion Species Manifesto: Dogs, People, and Significant Otherness* (Chicago: Prickly Paradigm Press, 2003), 35.

[5] Ibid., 45.

The goal of this dissertation will not be to resolve this debate regarding the differences or similarities of humans and other animals in its plurality of forms. Rather, I want to focus this discussion by engaging with a single point in this larger debate about humans and nonhuman animals (hereafter referred to as NHAs). There are debates about rationality to be sure, about emotional intelligence and dimensionality, about language and communication, and so on. But in few places has the debate been so quickly silenced, or has the widespread acceptance of our profound difference gone so largely unchallenged, as in conversations about morality – and more specifically, those regarding the ability to act morally.

We point to the lion who attacks and consumes the gazelle, and we – well, at least some of us – conclude that humans alone are capable of overriding instinct and of taking seriously the lives of others in the world (over and against our own interests at times). While there are many proposed divides between us and NHAs, perhaps none is as deep as this divide between the morality of humans and the amorality of NHAs. Perhaps corvids have self-awareness and rather exceptional rational capacities with respect to tool usage and problem-solving. But surely crows cannot be "good" or "bad" crows – crows who do "good" or "bad" things!

Let me pause to make two clarifications. First, I am not suggesting that many or even most humans think we have no moral obligations to NHAs. Rather, when I speak of NHA "amorality," I mean only to signal that we do not generally think of NHAs as acting morally or immorally themselves. To put it differently, we think of ethics as a uniquely human endeavor.

The second clarification concerns the labels "good" or "bad." To be sure, we sometimes employ these terms to praise or criticize NHAs, especially companion animals. When Fido sits upon command, we exclaim, "Good boy!" But when we use these terms in most circumstances, we are not suggesting that Fido's behavior (e.g., his sitting) is somehow *morally* praiseworthy. Even if Fido engages in an action for which we would assign specifically moral blame to another human (e.g., he mauls some unlucky passerby), we may blame his companion human, the passerby who antagonized him, or even the instincts associated with his breed. But again, I do not think we mean to commit ourselves to the idea that Fido is a "bad dog" in the sense of being a dog who has chosen to act in a morally blameworthy way.[6]

Transitioning from looser talk to explicitly philosophical discourse, this moral divide between humans and NHAs has been captured in familiar distinctions between moral agents and moral patients. Lori Gruen offers particularly direct definitions of the terms "moral agent" and "moral patient," as follows:

> Moral agents as persons have certain capacities that allow them to make reflective choices about their actions and to attend to those who may not be able to make such choices but who nonetheless have lives that will be affected, for better or worse, by [the agent's] actions. These latter non-persons are moral patients; they are the recipients of moral attention and concern, but they do not have the moral responsibilities that moral agents, as persons, do.[7]

[6] I should note here that Stephen Clark seems to disagree with my assessment here. He argues that a "good dog" does actually "show some signs of having preferred the paths of virtue to those of easy gratification" [1985, 50-51]. Presumably then, a bad dog does not show these signs or have these preferences. Still, my point above is only that in most instances involving NHAs we do not employ the labels "good" and "bad" in a moral sense of the terms, and that if we do, we do not mean to suggest that the NHA is an object of appropriate moral praise or blame.

[7] Lori Gruen, *Ethics and Animals: An Introduction* (Cambridge: Cambridge University Press, 2011), 60.

Now, to be clear, this distinction between agents/persons and patients/non-persons does not *necessarily* map perfectly onto humans and NHAs.

For example, as Gruen notes, the failure of particular humans to possess some feature deemed necessary for personhood (e.g., "intentionality, self-awareness, memory, imagination, a sense of existing over time") can be employed to locate such humans in the category of non-persons.[8] On the other hand, empirical proof that some feature deemed necessary for personhood is present in some particular NHA species (e.g., capacities for memory and self-conception in jays) can be used to argue that members of that species are in fact nonhuman persons.[9] This debate about personhood, however, is somewhat tangential to the present discussion, and I will not pursue it further.[10] My main purpose in bringing Gruen to the conversation has been simply to introduce the terms "moral agent" and "moral patient," as a means of beginning to explore what these terms signal about the moral status of (most) humans and (most) NHAs.

Still, while there have been notable exceptions[11], there has been a sustained resistance in the philosophical community to thinking of NHAs as agents. There are two significant reasons for this: the first being that most NHAs do not obviously possess the metacognitive ability that is necessary for evaluating or reflecting on one's choices, and

[8] Ibid., 64.

[9] Ibid., 59-60.

[10] To be clear, I am not suggesting that the debates about nonhuman personhood are completely unrelated. But this body of literature will not be the focus of my project. Moreover, I think arguments aimed at persuading readers that NHAs can be persons are misguided, at least insofar as the language of "personhood" seems to presuppose a hierarchy between humans and other animals. Additionally, I have worries about personhood debates too easily lending themselves to anthropomorphic readings of NHA behavior.

[11] E.g., Stephen Clark, Steven Sapontzis, Evelyn Pluhar, David DeGrazia, Paul Shapiro, Lawrence Johnson.

the second being that it does not make sense to reward or punish beings who lack the capacity for metacognition. Insofar as metacognitive ability and responsibility are built into the concept of the "moral agent," the exclusion of NHAs from the category seems correct.

But, with more and more empirical literature defending the presence of pro-social, altruistic, and other-regarding behaviors in other species, continuing to describe NHAs as mere moral patients seems to get the facts wrong. So, perhaps, the assumption that most NHAs are moral patients (following from the idea that most NHAs do not obviously possess whatever capacities are argued to be necessary for moral agency) is also problematic. That is, it seems that the space of moral patienthood does not adequately capture the moral lives of NHAs – even if most NHAs are not moral agents.

There is a recent, alternative account of the moral status of NHAs that I will explore and develop herein – a view defended by Mark Rowlands in his book *Can Animals Be Moral?* (2012). Rowlands recognizes that the assumed disjunction between moral agency and patienthood ignores crucially important conceptual space for thinking of a third kind of being in the moral sphere – namely, the "mere moral subject." He acknowledges that, for much of history, NHAs have been assigned the status of moral patients, granted that most philosophers have viewed them as lacking capacities required for the attribution of moral agency – metacognition, most importantly. But, in light of the behaviors we see some other species exhibit, Rowlands proposes that some NHAs may be better understood as moral subjects rather than as moral patients.

To be clear, the point of the present work is to take seriously the idea that there may be another way of acting for moral reasons (i.e., besides being a moral agent) and to explore what might follow if we grant that Rowlands's account is plausible. Rather than endorsing the historically entrenched view that agents alone can act morally, I take seriously the idea that mere moral subjects might also act morally. In short then, my goal here is not to defend all of Rowlands's assumptions or to endorse every move of his argument. Rather, since he offers a radical and interesting way of describing the morality of NHAs, I wish to explore the implications of his view and propose some answers to open questions.

The most significant of these open questions, and the one that is at the heart of this project, centers on exploring what capacity or mechanism makes possible moral action in NHAs. My proposal is that an appeal to empathy is a plausible response to this open question in Rowlands's view, and that the capacity for empathy can make possible this linking of an individual's perceptions to the morally salient features of the world.

Chapter II provides a detailed summary of Rowlands's book, focusing on both his account of a moral subject as well as his arguments for thinking that (some) NHAs are moral subjects.

Chapter III begins the exploration of the implications of Rowlands's view. It is here that I begin to articulate a response to what I have called the most significant open question in Rowlands's account (i.e., what capacity or mechanism does the work of linking perceptions to the morally salient features of the world). My response to this question brings the capacity for empathy front and center. Additionally, in this chapter, I introduce what I call the *adaptivity criterion*, building from my introduction of empathy into the discussion and Rowlands's account of the induction of NHAs into moral practices. Adaptivity, in this context, refers to the ability to respond to feedback about one's behaviors from the members of one's moral community.

Chapter IV begins the discussion of what the relevant meaning of "empathy" at work in my previous chapter is. After a brief survey of the early debates in the history of philosophy, I explore the three prominent theories of empathy on offer in contemporary theory of mind debates.

Chapter V provides a conclusion to the discussion of the previous chapter. Therein, I both endorse an interaction theory view of empathy, appealing primarily to direct perception but also to social competency, and engage with empirical literature about the capacity for empathy (so understood) in NHAs. The empirical literature engaged serves the role of undergirding the claim that NHAs from a variety of species possess the kind of empathy that I have argued is necessary for an NHA to be a mere moral subject.

Chapter VI will include an examination of some implications of accepting Rowlands's distinction between mere moral subjects and subjects who also act as agents.

And finally, *Chapter VII*, will include a few points of discussion, express one major worry about this model of mere moral subjecthood and its relationship to agency, and suggest some questions for future research.

Before moving forward, I should pause to acknowledge how indebted I am to Mark Rowlands in what follows. It is Rowlands's recent work on this subject that has motivated my dissertation. In my view, Rowlands makes a tremendous contribution in entering the debate regarding the moral status of NHAs by giving a theoretical account of how being motivated to act for moral reasons is sufficient for being a moral subject. I hope that the majority of what follows will be seen as a friendly exploration of and addendum to his view.

2

Outlining Rowlands's View

The object of this chapter is twofold: (1) to get clear on what is meant by "mere moral subject," differentiated from "moral patient" and "moral agent;" and (2) to recapitulate Rowlands's argument for thinking of NHAs as moral subjects as completely and charitably as possible. With that in mind, this chapter will begin with a broad discussion of the difference between Rowlands's three types of moral status, then move to a breakdown – chapter by chapter – of his book "Can Animals Be Moral?" (2012).

Initial definitions

As I mentioned in the introduction of the present work, Rowlands has recently offered a promising theoretical account of how we should conceive of the moral status of at least some NHAs. In outlining his terms more systematically, Rowlands distinguishes between moral patienthood, moral agency, and moral subjecthood. Moral patients are defined as follows: "X is a moral *patient* if and only if X is a legitimate object of moral concern: that is, roughly, X is something whose interests should be taken into account when decisions are made concerning it or which otherwise impact on it."[1]

In contrast, regarding moral agency, Rowlands says 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 understood) for, its motives and actions."[2] Moral agents are thus beings who are able to act for moral reasons and who can utilize moral scrutiny to

[1] Mark Rowlands, *Can Animals be Moral?* (New York: Oxford University Press, 2012), 74.

[2] Ibid., 75.

evaluate the relative goodness of those moral reasons. From this capacity for evaluation, it follows that a moral agent can be judged (i.e., praised or blamed) for both the motives she entertains and/or the actions she chooses.

These are familiar categories, and I do not think that Rowlands is employing the terms in any particularly controversial way. Moral subjecthood, however, is a new category that will ultimately require a great deal of unpacking. For now, we need only recognize that the moral subject occupies a conceptual space between the moral agent and the moral patient that has been ignored – that is, according to Rowlands, "X is a moral *subject* if and only if X is, at least sometimes, motivated to act by moral reasons."[3]

Before offering a detailed summary of Rowlands's account of a moral subject and his arguments for the inclusion of NHAs in the category of moral subjects, I offer a simplified comparison of our three types of moral status for quick reference (Table 1):

Table 1: Preliminary comparison of three categories of moral status

Category	Object of moral concern	Ability to be motivated to act by moral reasons	Ability to evaluate one's moral reasons or actions
Moral Agent	X	X	X
Moral Subject	X	X	
Moral Patient	X		

It is worth noting that this table is meant as an introductory gloss or easy point of reference. I should note immediately that this table might be misleading in two respects.

[3] Ibid., 89.

First, the table above is not meant to signal, strictly speaking, the existence of three different *kinds* of moral beings. We will see that Rowlands characterizes the difference between moral agents and moral subjects as a difference of degree – with moral patients being of a different kind.[4] To make this a bit clearer, moral subjects and agents are both beings who act reliably for moral reasons (though they do so in virtue of different tools or capacities). Since he characterizes moral patients as different in kind than moral subjects, it seems to follow that patients can be objects of moral concern but lack any capacity to act for moral reasons or to exercise moral scrutiny.[5]

Second, moral agency is not necessarily a static status in Rowlands's account. Rather, as will be seen, he leaves open the possibility that agency has to be evaluated not only from individual to individual, but also perhaps from particular action to particular action. For Rowlands, a subject's acting agentially has everything to do with whether or not she understands the moral facts of the matter at hand. Moral subjecthood and agency are then better understood as being on a spectrum rather than as being separate, static kinds in his view.

In light of these clarifications, it may be fairer to represent Rowlands's schema as follows (Figure 1), though some of the details from Table 1 are lost in doing so:

[4] Ibid., 241.

[5] My wording here is careful, and it betrays a worry I have about Rowlands's account that I will return to later. But to note my concern for now, it is not obvious to Rowlands's reader that he actually locates any beings in the sphere of moral patienthood exclusively – that is, that any beings occupy that space who are *not* also moral subjects. My worry is then that his concept of moral patienthood, while conceptually distinct from these other categories of moral status, proves to be undertheorized. There is a way of responding to my challenge, one that Gruen articulates – a sort of contextualized account of moral patienthood [Gruen (2011), 62]. For now, let me just flag the point that Rowlands's account of moral patients is somewhat lacking, in my view, and will require further attention.

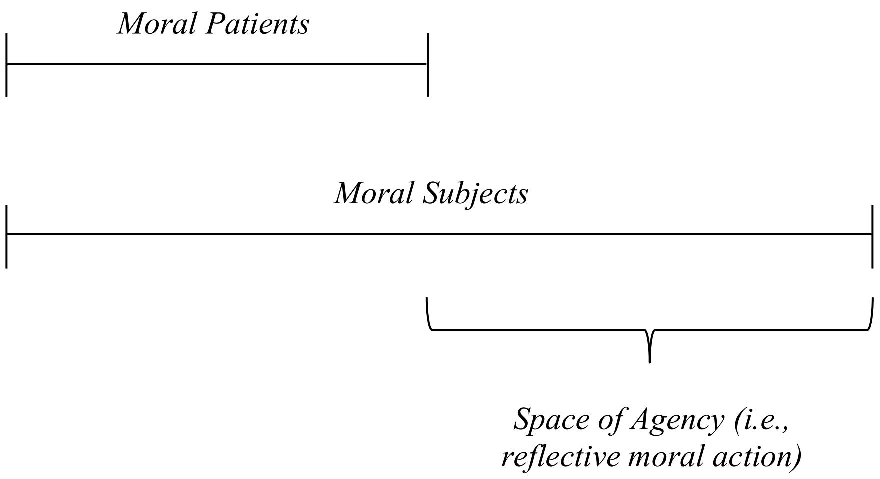

Figure 1: Revised comparison

Though a preliminary sketch to be sure, some idea of what Rowlands's means when referring to moral agency, subjecthood, and patienthood is now available. We can now turn to a much more in-depth examination of his argument.

Rowlands's Ch. 1 - "Animals, Emotions, and Moral Behavior"

Rowlands's first chapter opens with a presentation of a number of stories that are intended to support the overarching thesis of the book: namely, that NHAs can act morally. Rowlands understands such examples as evidence that NHAs demonstrate "concern," a term he intends to include a variety of emotions (e.g., compassions, sympathy, grief, courage.) But in any case, when the NHA in question demonstrates concern for another's wellbeing, regardless of the particular details, some particular emotion serves as the concerned NHA's motivation to act. Now, to be clear, one might be positively or negatively concerned about another's welfare, as Rowlands puts it – such that other emotions like malice, spite, and cruelty can fall under his broader label of concern. So whether one delights in another's good fortune, resents that other's good fortune, or actually wishes for and is pleased by ill fortune befalling that other, we can say that the experiencing being is concerned – positively or negatively – about the other.

Rowlands goes on to suggest that any evidence of concern for another is “the hallmark of moral attitude,”[6] such that all of his concern emotions fulfill this requirement and can serve as moral reasons. Moral reasons are, most simply, considerations that motivate a being to act one way or another with respect to some other being. Already, we know that Rowlands is going to argue that, insofar as NHAs evidence concern, we can say that they act for moral reasons.

Rowlands acknowledges that some reader may already be skeptical, but he takes it that there are enough stories that suggest that NHAs demonstrate concern (in either positive or negative emotions) to provide at least a prima facie case for assuming that NHAs might have emotional states with moral content. What is fascinating, in light of the multitude of anecdotes we have about NHA behavior that appears to show concern about another, is that philosophers have historically “reached a near unanimous decision”[7] that NHAs cannot be moral – that they cannot act for moral reasons. This view that NHAs cannot act for moral reasons (such that when they act from emotion, such emotions cannot be labeled as moral) has at its base “spurious” reasoning that amounts to a kind of “magical thinking.”[8] Much of the remainder of Rowlands’s book is dedicated to defending this analysis.

In short then, the thesis of Rowlands’s project is captured as follows: NHAs can act morally; when they do, they are motivated by moral emotions (“emotions that possess

[6] Rowlands (2012), 8.

[7] Ibid., 14.

[8] Ibid.

identifiable moral content") to act for moral reasons; and acting on the basis of a moral reason is a sufficient (but not necessary) condition for acting morally.[9] This last part of the thesis sets Rowlands apart both from sentimentalist accounts of moral action (e.g., those of Hume and – to some extent – Darwin) and from accounts of morality that demands the presence of more than so-called rudimentary forms of morality for genuine moral action to be possible (e.g., Hume, Darwin, de Waal). According to Rowlands's analysis, it is this distinction between rudimentary forms of morality and genuine ones that has undergirded the view that humans alone can act for moral reasons – because thinkers like Hume, Darwin, and de Waal have all required "increased cognitive sophistication" of some kind to be present for full moral action.[10]

At this point, Rowlands concludes his first chapter by outlining five specific claims he intends to defend: "(1) Animals can be moral *subjects*," "(2) To be a moral subject is to be *motivated* to act by moral *considerations*," "(3) In the case of animals, these considerations take the form of *morally laden emotions* (or *moral emotions*)," "(4) An emotion is morally laden if and only if it involves (in a sense to be clarified) a moral *evaluation* or *judgment*," and "(5) Morally laden emotions motivate animals to act by providing *reasons* for those actions."[11]

[9] Ibid., 15.

[10] Rowlands elaborates on this claim, as follows: Hume required that one develop some virtues that seem cognitively-demanding (e.g., justice), Darwin required that the ability to approve or disapprove of one's motive for actions be present, and de Waal requires that there be an explicit understanding of the value of community [see pp. 31-32].

[11] Rowlands (2012), 33-35.

Rowlands's Ch. 2 - "Attributing Emotions to Animals"

The primary purpose of Rowlands's second chapter is to clarify what he means by the term "emotion" and to address some challenges to attributing emotions to NHAs. Broadly, Rowlands understands emotion as involving intentional states – as "states that possess intentional or propositional content."[12] It is in this definition that we see the first challenge to attributing emotion, so understood, to NHAs. Immediately, certain philosophers (e.g., John Deigh, Donald Davidson, Stephen Stich) spring to mind, as they argue that the having of emotions – understood as involving the entertaining of propositional content – requires language.[13]

In responding to this challenge, Rowlands draws a distinction between moods, which may not be about any thing in particular, and emotions. Emotions, on the other hand, are about something – they have "reason-giving, justification-specifying" content.[14] Rowlands understands the problem of attributing emotion to NHAs, as articulated by Davidson and Stich, in terms of the question of what sortals NHAs would use (since we humans do not know how NHAs represent objects in the world). But this is an empirical question, according to Rowlands. We have good reasons to assume, in the meantime, that some NHAs possess concepts, concepts that admit of some *de dicto* content-ascription – even if for the time-being we can only be sure of the *de re* ascriptions of NHA concepts.[15] Or, put differently, "[t]o suppose that there is no content available for us in *de dicto*

[12] Ibid., 40.

[13] Ibid.

[14] Ibid., 42.

[15] In fact, Rowlands suggests that Colin Allen has already outlined conditions under which an NHA can be said to possess a concept [see pp. 54-57].

ascriptions is to suppose animals do not represent the world…[which is] a breathtakingly implausible assumption."[16] But setting aside the barriers to our knowledge of how NHAs represent objects in the world, Rowlands will go a step further to contend that we can still legitimately employ *de dicto* ascriptions of concepts to describe NHA behavior – in order to "track," or more loosely to make a "best guess," as to the propositional content of NHA concepts.[17]

Consider the famous case of the dog eyeing a squirrel who has just run up a tree. We might say that the dog believes the squirrel to be in the tree, but of course, this may not in fact be what the dog believes. Perhaps the dog only distinguishes between chaseable and non-chaseable things in the world, such that the squirrel simply falls into the first of these categories with no further refinement of the concept being utilized by the dog. Rowlands proposes the following solution to the problem of not being able to access the propositional contents of NHA concepts: we need to "relativize content-ascriptions to contexts."[18]

For example, some proposition may be anchored to the human context, or alternatively to the canine context. We can say then that the proposition in the human context tracks the proposition in the canine context so long as there is a "reliable asymmetric dependence between the concept expressed by 'squirrel' (as anchored to a human) and the concept expressed by 'chaseable thing' (as anchored to a dog...)."[19] Or

[16] Rowlands (2012), 54.

[17] Ibid., 58.

[18] Ibid., 59.

[19] Ibid., 60.

put differently, the truth of [H: p] guarantees the truth of [C: p*], such that though the dog cannot even entertain [H: p], we humans can track [C: p*]. According to Rowlands's view, we can offer a plausible explanation of the behavior of NHAs (or even kinds of humans with whom we cannot communicate verbally) when the follow three conditions obtain: "(1) the ascription of *p* is the *de dicto* version of a *de re* ascription that is correct;" "(2) There exists a proposition *p** that the subject does entertain and the truth of [H: *p*] (that is, *p* anchored to its context) guarantees the truth of [C: *p**] (that is, *p** anchored to its context);" and "(3) *p* guarantees the truth of *p** by way of a reliable asymmetric connection between the concepts expressed in the subject terms of [H: *p*] and [C: *p**]."[20] Again, the point is that we have a way of describing NHA behavior reliably, from within our own human context, despite our not having access to all of the details of some propositional content of a concept belonging to an NHA – even when the NHA in question cannot even entertain the same content that we can.

Returning then to the contents of emotions more specifically, Rowlands suggests that emotions contain both factual and evaluative contents. So, to use his example, we know of the ancient mariner that he has a skinny hand and that he is the sort of thing to be feared. Moral emotions – or what we have been calling morally laden emotions – are then simply those that have a "moral hue" to their evaluative component.[21] Again borrowing Rowlands's example, when Smith is snubbed by Jones, Smith's emotion includes both the factual component that the snubbing occurred as well as the evaluation that Smith deserved better treatment from Jones. But notice that all that is required for an emotion to

[20] Ibid., 63.

[21] Ibid., 65.

track a proposition is that a true proposition makes sense of the emotion, not that the experiencer of the emotion entertains (or be able to entertain) the evaluative proposition. Hence, just as Rowlands described the manner in which we can defensibly attribute intentional states to NHAs, he now extends this view to attributing morally laden emotions to them.[22] In short, the purpose of this second chapter is to offer an account of how "the behavior of animals may, legitimately, be explained by appeal to states that possess both factual and evaluative content."[23]

Rowlands's Ch. 3 - "Moral Agents, Patients, and Subjects"

With Rowlands now having laid the groundwork for how we can attribute morally laden emotions to NHAs, even if they cannot entertain the propositions that undergird such emotions, he moves to draw out the distinctions between moral agents, subjects, and patients. The contrast between the first two will be of the highest import for his argument, and that contrast will turn out to mirror the distinction between motivation and evaluation.

First then, in the context of the larger goal of this chapter, it is important for Rowlands to address those who have made arguments that NHAs are mere moral patients. Leaders in the animal rights and animal welfare camps, namely Regan and Singer respectively, have located NHAs here. And there is something about this view that seems consistent with common sense. Though we may be inconsistent in recognizing the moral claims NHAs demand of us, as Rowlands puts it, we clearly have reservations about taking a chainsaw to a living dog. He takes it to be uncontroversial that NHAs

[22] For the detailed conditions under which an emotion can be said to be morally laden, see p. 69.

[23] Rowlands (2012), 70.

make at least some claims on us (i.e., as moral patients).[24] Understanding NHAs as moral patients coincides with many of our intuitions. And if we understand the only other option to be the category of the moral agent, the picture becomes clear as to why even some of the most vocal advocates for NHAs have not advocated for their reclassification.

Rowlands does acknowledge a few philosophical outliers who have done work to suggest that NHAs could be considered as moral agents (e.g., Stephen Clark, Steven Sapontzis, Evelyn Pluhar, David DeGrazia).[25] Though these philosophers have utilized different kinds of arguments in order to support the notion that some NHAs might be moral agents, Rowlands finds all varieties of such an argument to face the same implausible consequence: namely, that if NHAs are moral agents, then they have moral responsibility for their actions. It was in fact this kind of thinking that led to the seemingly absurd practice of trying and sentencing NHAs for perceived crimes at one point in history; Rowlands "assume[s] few would wish to recommend a return to this practice."[26] So, in keeping with the majority view, Rowlands will ultimately agree with the claim that NHAs are not moral agents, while holding that some NHAs can however be motivated to act by moral reasons – such that the idea of moral patienthood mischaracterizes the moral status of some NHAs. In short, he takes it that conceptual space lies open between these two familiar categories of agency and patienthood that may better characterize the nature of NHA morality: namely, the space occupied by the moral subject.

[24] Ibid., 74-75.

[25] Ibid., 75-84.

[26] Ibid., 83.

As Rowlands puts it, “[t]he concept of a moral subject has almost invariably been conflated with that of a moral agent: to say that X is motivated to act by moral considerations is almost invariably thought to be equivalent to the claim that X is responsible for, and so can be praised or blamed for, what it does.”[27] Notice how this conflation between subject and agent entails another conflation between motivation and responsibility. It seems clear, though, that these are distinct concepts. One’s having a motive to act and one’s being responsible for (in the sense of being praised or blamed for) an action are different. Rowlands asks us to imagine that his wife has hypnotized him, such that whenever she utters the word “rosebud,” he desires to mop the floor. Rowlands argues that, under such conditions, his mopping of the floor would not be praiseworthy – “it is the result of a motivational state that is outside of [his] control.”[28] After arguing that his mopping the floor is both a reason and his reason (rather than an ownerless one), Rowlands contends that we have good reason to think that the motivation behind an individual’s action and an evaluation of that action are separable – at least in the general sense.

But he notes a powerful worry: namely, that this separation of motivation and evaluation may not hold in the specifically moral realm. In fact, he admits that “there are persuasive arguments in support of the claim that being a moral agent is both a necessary and sufficient condition for being a moral subject” because “motivational states cannot make a normative claim on their subjects unless those subjects have control over those

[27] Ibid., 89.

[28] Ibid., 90.

motivations."[29] So it is this bond, this "unquestioned connection,"[30] between normativity and control that Rowlands is going to undermine, in his effort to defend the view that some moral subjects are not moral agents (though all moral agents are subjects).

Rowlands's Ch. 4 - "The Reflection Condition: Aristotle and Kant"

The assumed and seemingly unbreakable link between normativity and control has deep roots in the history of philosophy. Citing Aristotle's account of virtue as outlined in the *Nicomachean Ethics*, Rowlands reminds us of Aristotle's joint requirements that the agent know she is engaging in a virtuous action, choose to perform that action precisely because it is virtuous, and make this decision from a stable disposition or state. And since NHAs probably cannot understand what counts as a virtue (or choose to perform some act because it is virtuous), they fall outside the parameters of moral beings according to Aristotle.

Rowlands refers to this Aristotelian line of thinking as 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."[31] Moral agents meeting this reflection condition is imperative if we are to hold them responsible – to praise or to blame them – for their actions. Virtues, it seems, cannot be "a matter of nature" but instead a product of "our efforts" if we are to evaluate actions morally.[32] So for the Aristotelian, sentiments only exert normative grip

[29] Ibid., 97.

[30] Ibid.

[31] Ibid., 102.

[32] Ibid., 105.

on a being insofar as she demonstrates "control over whether she has this sentiment" or "control over whether she acts on it or does not."[33] Or put more sharply, "…without control there can be no normativity."[34]

Similarly, Immanuel Kant (on Christine Korsgaard's interpretation) holds a related view, according to Rowlands. At the heart of Kantian morality is the ability to understand, reflect upon, and either endorse or reject principles "on which we are inclined to act."[35] For the Kantian, actions or emotions that cannot be objects of "normative self-scrutiny" are simply not moral phenomena.[36] Just as with Aristotle, the notion of control plays a necessary role in making sense of normativity.

Rowlands's fourth chapter then wraps up with a brief discussion of what the reflection condition, adopted in both Aristotelian and Kantian accounts of ethics, means for NHAs. Since NHAs have no control over their motivations, those motivations are denied normative status. And since their motivations are understood as non-normative, the history of philosophy has concluded that NHAs cannot be moral subjects (because, up to this point, there has been a persistent conflation between moral subjecthood and moral agency). All of Rowlands's remaining chapters attempt to undermine this orthodox view and to consider implications of the newly defended view (i.e., that some moral subjects are not agents).

[33] Ibid., 108-109.

[34] Ibid., 109.

[35] Ibid., 110.

[36] Ibid., 111.

Rowlands's Ch. 5 - "The Idiot"

Chapter 5 of Rowlands's book details a thought experiment involving a character named Myshkin. Myshkin engages in many seemingly good acts, evidencing kindness and compassion. His motivations for these acts are emotions, sentiments that join together in what Rowlands will again call "concern." Though he experiences sentiments that motivate him to act in the interests of or on the behalf of others, he cannot scrutinize those sentiments before acting – that is, he lacks critical moral scrutiny. From the orthodox view of morality (i.e., that which includes the reflection condition), Myshkin's motivations and actions are not moral ones.

One challenge in proving that Myshkin is a moral subject is to transition from the claim that Myshkin's motivations/actions seem good to the claim that they actually are good. To begin this leg of the argument, Rowlands introduces another character: Marlow. Marlow represents the ideal moral judge, the being who is always right in response to questions of the moral variety – that is, he is a being with exceptionally accurate moral scrutiny. Marlow, knowing nothing of Myshkin, independently concludes that there are morally correct sentiments, and that "for all identified circumstances, [those of Myshkin] are the morally correct sentiments to have and the morally correct actions to perform."[37] So while we cannot appeal to *phronesis* to explain how Myshkin gets things right, when he does, Rowlands posits that Myshkin nevertheless has a kind of moral "sensitivity" – that is, he is "sensitive to some of the features of a situation that make it a good one (for

[37] Ibid., 128.

example, the happiness of others), and to some of the features that make it a bad one (for example, the suffering of others)."[38]

Myshkin's moral sensitivity is limited. He can only be sensitive to some of the morally salient features of any situation (rather than additional features that would require *phronesis* to recognize), and he will inevitably make moral mistakes. Nevertheless, Rowlands argues that we can attribute to Myshkin the possession of a kind of "moral module."[39] After toying with a number of formulations, the final account of Myshkin upon which Rowlands settles is as follows:

> (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 scrutinizes the *deliverances* of this module.[40]

In contrast to Myshkin, a final account of Marlow is also provided, one that will prove critical for the remainder of Rowlands's argument:

> "(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,

[38] Ibid., 131.

[39] Note that the term "module" here is not meant to entail any particular theory in cognitive science; rather, the term here only stands in for the kind of cognitively impenetrable mechanism that Rowlands has in mind [see pp. 146-147].

[40] Rowlands (2012), 146.

> therefore, (5) is able to engage in effective critical scrutiny of the deliverance of this module."[41]

Though these characters are meant to serve a conceptual point (rather than to be "psychologically realistic"[42]), they serve a critical function in highlighting what exactly has traditionally separated moral subjects from moral patients: namely, the difference between the fourth and fifth elements of Marlow's and Myshkin's characterizations, respectively. These characters will then be employed in the service of the next leg of the argument: that "[t]he related ideas of access and scrutiny…have little substance and cannot do the work required of them [i.e., they cannot 'give a subject control over her motivations']."[43]

Rowlands's Ch. 6 - "The Phenomenology of Moral Motivation"

Referring to the character of Marlow introduced in the previous chapter, Rowlands begins this chapter by describing Marlow's attainment of the status of moral subject as following the "ASCNM route" – that is, moving from access, to scrutiny, to control, to normative status, and finally to moral status.[44] To elaborate, according to the orthodox view attributed to thinkers like Aristotle and Kant, for beings (e.g., Marlow) to have the moral status of a moral subject, one must have motivations that have normative status. For motivations to have normative status, one must have control over those motivations. To have such control, one must have the ability to engage in critical moral

[41] Ibid., 148-149.

[42] Ibid., 149.

[43] Ibid., 151.

[44] Ibid., 152.

scrutiny. And to exercise this scrutiny of one's motivations, one must have access to one's process(es) of moral deliberation.

Though Myshkin possesses a moral module mechanism – such that his perceptions, emotions, and actions are linked in such a way as to provide reliable sensitivity to some of the morally salient features of situations, he is still cast out from the category of moral subject due to his inability to proceed down the ASCNM route. But this ASCNM schema is terribly problematic, as Rowlands argues. Immediately, he sees an issue with linking moral scrutiny of one's motivations to having control over those motivations.

He pauses to make a few clarifications. First, by "moral scrutiny," he has only a general concept in mind: namely, "the ability to compare one's motivations with antecedently accepted principles…and to ascertain their mutual compatibility or incompatibility."[45] Second, he draws a distinction between *mere* critical scrutiny (such that the scrutiny is ineffective and leads to moral mistakes about one's motivations) and *effective* critical scrutiny (such that critical scrutiny entails some level of success). He suggests that we treat Marlow as having effective critical scrutiny, for the present purposes. Third, Rowlands distinguishes between the *ability* to engage in effective critical scrutiny and the *exercise* of that ability.

Notably, even those who hold the orthodox view do not think that moral subjects[46] must demonstrate effective moral scrutiny with respect to every motivation in every situation. As Rowlands puts it, if this ability was expected to be exercised on every

[45] Ibid., 157.

[46] Again, all moral subjects are agents in the orthodox view; the distinction collapses.

occasion, this necessary condition for moral subjecthood “would almost certainly entail that there are no moral subjects.”[47] So, instead, it seems that the *ability* to scrutinize is what serves as a necessary condition for being moral subject. With these clarifications in mind, Rowlands indicates that individuals like Marlow who engage in critical moral scrutiny are often called “morally autonomous” subjects. It is, in this language, “the *significance* of this autonomy [i.e., that autonomy entails control of one’s motivations]” that Rowlands wishes to challenge.[48]

The present question is framed around what properties of scrutiny link to (i.e., yield) control. The remainder of his chapter is dedicated to rejecting one kind of answer that might be given. One might contend that the phenomenological properties of moral deliberation play the linking role between scrutiny and control, but Rowlands finds this response to be lacking. He considers the case of a super-blindsighted subject. In such a case, it does *not* seem plausible to argue that the subject cannot see – if by “see,” we mean to indicate that she is a visual subject but not a subject with visual consciousness. She still relies on her eyes (and their deliverances) to provide data that is processed in the visual cortex of her brain, according to Rowlands.

He explains thusly: “It is just that these processes that usually result in a perception possessing visual phenomenology now result in a feeling of… a more visceral character.” He goes on: “Rather than denying that she is seeing, [Rowlands] think[s] it is far more plausible to claim that this is what seeing consists in for the super-blindsighted

[47] Rowlands (2012), 158.

[48] Ibid., 160.

subject."[49] Myshkin is, essentially, an individual with "moral super-blindsight."[50] Just as *seeing* admits of "differing visual phenomenologies," *being moral* may well admit of "differing moral phenomenologies."[51] The point of this discussion is to eliminate one candidate (i.e., the phenomenological one) for responding to the question of how scrutiny and control are linked, on the grounds that there may be different moral phenomenologies, such that we would be incorrect in denying to some individual "the status of moral subject simply because her associated moral phenomenology differs from ours."[52]

Rowlands's Ch. 7 - "Moral Motivation and Metacognition"

In this chapter, Rowlands returns to the same question from the previous chapter: what properties of moral scrutiny yield a subject control over her motivations? Having dismissed the phenomenological reply, Rowlands now turns his attention to the metacognitive response. While Marlow can reflect on his motivations, Myshkin is "at the mercy" of his motivations.[53] Perhaps this is enough of a reason to conclude that Myshkin's motivations are merely *causes* that lack any normative dimension, while the motivations of Marlow constitute *moral reasons*.

Rowlands contends that to assume Marlow's metacognition provides or confers control over his motivations is to commit the fallacy of the "miracle-of-the-meta."[54] By

[49] Ibid., 166.

[50] Ibid., 167.

[51] Ibid.

[52] Ibid.

[53] Ibid., 170.

[54] Ibid., 171.

drawing an analogy to a criticism of the HOT model of consciousness, Rowlands suggests the following: "Whether disposition or occurrent, intransitively unconscious thoughts do not make us transitively conscious of what they are about."[55] In Rowlands's assessment, the idea that metacognition could account for the normative status of a motivation is a similar kind of "miraculous thinking."[56] Just as Myshkin is "at the mercy" of his sentiments, Marlow is "at the mercy" of his metacognition. Marlow simply cannot actually be aware of (or conscious of, if you prefer) all of the features of a situation in which he finds himself and over which he lacks control. Or at least, Marlow is "hostage to empirical fortune" regarding what situations he ends up in (and subsequently what degree he can be conscious of the variety of contextual features of any situation).[57] It remains entirely unclear why we should assume that metacognition could yield control of one's motivations.

In short then, the work of this chapter and the previous one has been to demonstrate that there is no "satisfactory account of the connection between effective critical scrutiny and control."[58] The grounds upon which the ASCNM schema makes a case against the moral subjecthood of beings like Myshkin (who stands "proxy for other animals"[59]) are ultimately "unsubstantiated" and "seemingly mysterious,"[60] granted that

[55] Ibid., 177.

[56] Ibid., 178.

[57] Ibid., 184-185.

[58] Ibid., 188.

[59] Ibid., 189.

[60] Ibid.

candidates for establishing the link between moral scrutiny and control are shown to be wanting.

Rowlands's Ch. 8 - "Moral Reasons and Practice"

The next part of Rowlands's argument for the moral subjecthood of some NHAs centers on his response to the "moral practice hypothesis" – "[t]o be a moral subject it is necessary that one belong to a *moral practice*."[61] According to Rowlands, thinkers like Ludwig Wittgenstein and Wilfrid Sellars differ from other thinkers who he has previously considered insofar as they will not locate the source of normativity as internal to an individual. Rather, the view is that "[n]ormativity is essentially embedded in social practices."[62] Rowlands interprets this broad view, understood also as an objection to the moral subjecthood of NHAs, in two distinct ways: (1) as the practice hypothesis – "one might argue that an individual's belonging to a practice is a necessary condition of that individual's possessing *reasons* – of any sort;" and (2) as the moral practice hypothesis – "one might argue that while belong to a practice is not required for the possession of reasons as such, it is a necessary condition of the possession of specifically *moral* reasons."[63]

The first of these approaches is circular. The practice hypothesis is intended to "groun[d] the possibility of meaning or content," with one implication being that NHAs are denied the ability to possess reasons because they do not belong to practices.[64]

[61] Ibid., 190.

[62] Ibid., 191.

[63] Ibid., 192.

[64] Ibid., 197.

Rowlands gives the example that he is both patting his head and rubbing his stomach simultaneously. But does this example evidence one or two intentions? Should it be interpreted as one action or two actions then? As Rowlands sees it, practice is a conjunction of actions, any action is individuated by the particular intention behind it, and the content of an intention individuates that particular intention. This first formulation of the view, as sketched in the previous paragraph, "cannot explain how content is possible for the simple reason that it presupposes content."[65] There is no justification provided in this formulation for thinking that content cannot exist apart from practice. Moreover, even if the practice hypothesis is granted, one could defend the idea that NHAs can and do belong to practices of the relevant sort. Rowlands cites the examples of vervet monkeys, ring-tailed lemurs, capuchins, Diana monkeys, and Campbell's monkeys, all of whom utilize variations in alarm calls to communicate with conspecifics regarding the type of predator who is approaching. Regardless of what such cases do or do not suggest about the cognitive abilities of these NHAs, Rowlands takes it as fairly obvious that they constitute a "communicatory *practice*."[66]

Turning then to the second of these approaches, recall that this formulation makes a claim about specifically moral reasons (rather than all reasons). Subsequently, the bulk of the present summary of this chapter will focus on this second formulation, as it is the more challenging and more relevant one with which to contend. Rowlands appeals to Dixon as a proponent of the view that the moral practice hypothesis excludes NHAs from the status of moral subjects. For Dixon, again according to Rowlands, the idea is that human children are initiated into moral practices via question-asking that demands of

[65] Ibid., 198.

[66] Ibid., 199-201.

them that they articulate reasons for what they do. It is this kind of moral training that differentiates humans from NHAs. Or to actually quote Dixon, "…the dog [who refrains from biting children] does not possess an understanding of even very simple moral concerns that prohibit such an action." She goes on: "…trained domestic animals do not participate in the *practice of morality*."[67] Rowlands interprets this as an obvious reiteration of the reflection condition he first addressed in Ch. 4. As such, he writes of Dixon that she "simply elaborates an idea that is fundamentally misguided [i.e., subject to the criticisms he has already issued against the reflection condition and the ASCNM schema's assumption of a link between scrutiny and control]."[68]

But Rowlands will add an additional worry about accounts like Dixon's: namely, that they mischaracterize what constitutes a moral practice. As he puts it, it is "undoubtedly true" that NHAs cannot give reasons (to themselves or others) with regard to their behaviors/actions, but Rowlands "argue[s] that there is little reason for supposing that this is the only way of understanding the idea of a moral practice."[69] Now, Rowlands and Dixon do agree on one point – that NHAs are not beings who are morally responsible. But staying true to his overall thesis, this simply indicates that NHAs are not agents; it remains plausible that NHAs fit within the broader category of the moral subject.

To begin to see what the moral practices of NHAs look like, and how they come to be inducted into them, Rowlands asks us to consider the case of his companion animal Hugo. Hugo is a German shepherd, trained to work with a bite sleeve. Rowlands remarks

[67] Qtd. in Rowlands (2012), 204.

[68] Rowlands (2012), 108.

[69] Ibid., 209.

on the fact that Hugo will enthusiastically engage and sometimes even tackle the wearer of the bite sleeve – when that wearer is Rowlands himself. But notably, when Rowlands's young son puts on the same bite sleeve, Hugo's behavior shifts dramatically to include a slow approach followed by a gentle nibble at the sleeve. As Rowlands puts it, Hugo's learning not to bite under certain circumstances is a "surface feature" of his larger profile – of his "form of life."[70]

This form of life follows from his being inducted into a moral practice, such that (among other traits) Hugo suppresses his own desires in certain situations. And the process of this induction is not the question-asking and reason-giving that Dixon has in mind. Rather, "…[that] process is a superstructure that is erected on a foundation of example."[71] Put differently, Hugo learns through example – the example of Rowlands and the examples afforded to Hugo by others as well. Rowlands, when thinking of a form of life, characterizes Hugo as having learned to value Rowlands's son, because of the repetition of Rowlands's behaviors towards his son that "impress" themselves upon Hugo. Crucially, "[i]f [the son's] tears were met instead with callousness or indifference [by Rowlands], then Hugo would be inducted into a very different type of moral practice, and so become a very different dog."[72]

Because Hugo demonstrates the capacity to act for moral reasons when he shows concern towards Rowlands's child (e.g., by behaving in ways that go against Hugo's own desire to bite down on the bite glove), the suggested conclusion is that Hugo is in fact a

[70] Ibid., 211.

[71] Ibid., 212.

[72] Ibid.

moral subject – a moral subject, who in the way of all moral subjects, first (or perhaps only) learns "by way of the deed."[73] The question of Hugo's moral agency is separate and depends on how one responds to the question of whether Hugo understands that hurting the boy would be wrong. But in any case, Hugo's status as a moral subject is clearly defended.

Rowlands's Ch. 9 - "Reconstructing Normativity and Agency"

As the title of this chapter indicates, it is now time for Rowlands to offer his positive view in full. Though elements of it are already clear and have been defended in previous sections, he accepts that "two puzzles" remain for which he will have to provide positive accounts in response: the first of which involves normativity and the second agency.[74]

Rowlands's work thus far has been negative, in the sense that his argument has been that "there are no persuasive reasons for supposing" that beings like Myshkin are not moral subjects.[75] But the need for positive work – reconstructive work – becomes obvious when Rowlands notes a possible implication of his negative argument to this point. Since the entire link between control and scrutiny has been undermined, setting aside the implications for NHAs for a moment, it is not clear how humans are moral subjects. After all, the orthodox view has been that moral motivation presupposes control; but if Rowlands is right in calling the "requisite notion of control…empty," then he has foreclosed the almost universally accepted account of moral action without defending

[73] Ibid., 213.

[74] Ibid., 214-215.

[75] Ibid., 217.

some other view as to how moral action is possible – even in the human case.[76] Crucially, Rowlands's positive story will have to be one that caches out both normativity and agency "in ways that do not rely on the problematic concept of control."[77] With this in mind, his response will be consequentialist and externalist in nature.

He asks us to consider two distinct ways in which we employ "ought:" there is the *moral* kind of "ought" that has been in the background of this entire dialectic, but there is also the *prudential* "ought."[78] So to parallel the characters of Myshkin and Marlow, Rowlands now introduces Prudence and Jude, respectively. Prudence has motivations in the forms of both cognitive and affective states, and these motivations compel her to act in ways that most would consider prudent. But she is unable to subject her motivations to scrutiny – that is, she lacks the metacognitive abilities to do so. Jude serves as our ideal prudential spectator, who endorses the claims that Prudence has prudent desires, that she has some means of effectively realizing those desires, and that the actions that follow are rightly understood as prudent actions.

The question is then whether or not we are to accept that Prudence is a prudential subject – that is, "an individual who acts for prudential reasons."[79] Rowlands suggests that to deny the prudential subjecthood of Prudence would be highly implausible and not tempting. Put differently, "[i]f one performs prudential action and one does so on the

[76] Ibid., 215.

[77] Ibid., 216.

[78] Ibid., 217.

[79] Ibid., 219.

basis of reliably prudential motives, that is sufficient for one to be a prudential subject."[80] A significant observation follows: "[t]he principle that *ought implies can* simply does not figure when we are dealing with the 'ought' of prudence."[81] The promise of this observation is that there must be some way of getting at what one *ought* to do that is meaningful even if one is at the mercy of one's motivations. And perhaps if that way can be identified with respect to prudence, an account of moral action can be offered that mirrors it.

Looking then at the prudential sense of "ought," Rowlands suggests that the normative grip of prudential motives requires an externalist account – such that "a given circumstance might possess certain prudentially salient features: features that impact, for good or for ill, the well-being of the subject."[82] When can they say of a motivation that it is prudent when it "tracks the good (in the sense of useful) features of the situation and thus allows the subject to make use of them in furthering his or her ends."[83] In such an externalist account, there is no need to invoke the language of control at all. Instead, we make sense of a motivation as a specifically prudential one by looking at only two things: the features of the situation and the degree to which the subject's motivations "cohere with the prudentially salient features of [that] situation."[84]

[80] Ibid.

[81] Ibid.

[82] Ibid., 220.

[83] Ibid.

[84] Ibid.

A parallel account of the moral sense of "ought" can now be generated.[85] Myshkin is motivated to act by a variety of affective states – and not just any affective states. He is motivated to act by morally laden emotions: that is, those emotions that are other-regarding and to which as a whole Rowlands has given the label of concern. And in light of his work to this point, "…there is no reason for supposing that these are not genuinely moral motivations rather than nonmoral facsimiles."[86]

The first step in reconstructing normativity, the first of his two puzzles, is to accept the specific theory of "objective consequentialism."[87] He offers the following definition of this position: "Objective consequentialism… defines a good or right action as one that actually produces good consequences – and the agent's expectations in performing the action are irrelevant to whether the action is right or wrong."[88]

The second step in reconstructing normativity will be to accept also the broader theory, under which subjective consequentialism is a species: namely, "evaluational externalism."[89] The relevant sense of externalism is so defined: "…the view that the

[85] Rowlands is forthright about the fact that Julia Driver, in her book *Uneasy Virtue* (2011), has also defended a view of moral action – more specifically, moral virtue – that is consequentialist and externalist. But he insists that his account is different in crucial respects from the one on offer from Driver, describing his own view as "somewhat more radical." For a more detailed discussion of the differences between their views [see pp. 221-222].

[86] Rowlands (2012), 222.

[87] Rowlands juxtaposes *objective* consequentialism against *subjective* consequentialism, with the latter version of consequentialist ethical theory contending that "…the goodness or rightness of an action is a function of subjective states of the agent who performs it" (e.g., expectabilism, which holds that some action is right *iff* the agent "expects that the consequences of the action will be good") [see p. 222].

[88] Rowlands (2012), 222.

[89] The broader theory (under which the subjective species of consequentialism falls) with which *externalism* contends is evaluational *internalism*: "…the moral quality of a person's action

moral quality of a person's actions or motives is determined, at least in part, by factors external to the person's agency."[90]

Tying this back to Myshkin, his so-called "moral module" is characterized by both its cognitive impenetrability and its sensitivity to some of the good- and bad-making features of the world. The status of a feature's goodness or badness, on this account, has nothing to do with Myshkin's internal, subjective states. In response to the question of what it is that does ground (objectively) the goodness or badness of some feature of a situation, Rowlands remains agnostic. In short, any objective consequentialist theory will do. As to who gets objective consequentialism right, so to speak, Rowlands means to leave his account open to be filled in by a range of plausible views, from those of the "hedonic utilitarian" to a view like Martha Nussbaum's "capabilities approach."[91]

Rowlands states explicitly that he is only committed to the following: (1) there are objective moral facts that make features of situations good or bad; (2) a moral subject need only be capable of detecting some of those features (e.g., feeling happiness in the presence of another's joy or sad in response to suffering); and (3) the subject's experiences follow from the detection of those morally-salient features (as opposed to those experiences constituting the good- or bad-making features of the situation).[92]

There is a brief aside here, responding to the worry that this account makes selfish action morally good. Perhaps Myshkin, in experiencing emotional contagion, alleviates

or motives – whether they are good or bad, right or wrong – is determined solely by factors internal to a person's agency" [see pp. 222-223].

[90] Rowlands (2012), 222-223.

[91] Ibid., 223.

[92] Ibid., 224-225.

his own discomfort in responding to the discomfort of another. This worry is quickly nipped. Rowlands contends that the selfishness of an action is a *not* "a function of its motivation [but of] its content."[93] So, to paraphrase the example given here, what matters in assessing whether Saint X is acting selfishly or not in helping other people – something he clearly wishes to do – is determined by the content of what Saint X desires (e.g., to get into heaven or to promote the happiness of someone else). As Rowlands puts it, "…the mere fact that a person always does what he wants – if it is a fact – does not mean that he is acting selfishly."[94]

Again, what makes Myshkin's motivations (that, for beings like him, only take the form of emotions) good or bad is determined by the extent to which his emotional responses are the objectively correct or incorrect ones with respect to the situation in which he finds himself. "It is this possibility of accord or discord," writes Rowlands, "that underwrites the normative status of Myshkin's motivations."[95] So long as Myshkin (or any other being) has a "reliable"[96] normative sensitivity to the morally salient features of the world, he can be said to be – and in fact is – a moral subject. At long last, this work culminates in a formal definition of a "minimal moral subject," as follows: "X is a *moral subject* if X possesses (1) a sensitivity to the good- or bad-making features of situations,

[93] Ibid., 226.

[94] Ibid., 226-227.

[95] Ibid., 228.

[96] This is Rowlands's term, but he uses it loosely to refer to any "significant proportion of the time" – such that "[t]here is almost certainly no precise line here [with respect to what percentage counts as 'reliable']" [see p. 229].

where (2) this sensitivity can be normatively assessed, and (3) is grounded in the operations of a reliable mechanism (a ‘moral module’).”[97]

This brings us to the second puzzle: the one regarding agency. Rowlands now needs to offer an account of how it is that humans have moral agency – when we do. But this second reconstruction, like the reconstruction of normativity before it, will also have to be made without appeal to the concept of control. Instead, Rowlands will appeal to the concept of “understanding.”[98]

Remember that Marlow has been described as like Myshkin in some ways, but as different from him in two respects: Marlow’s capacities for access and scrutiny.[99] Because of these metacognitive abilities, Marlow’s differences from Myshkin can be further fleshed out as follows: (1) Marlow can make “fine-grained distinctions between types of emotional states [e.g., between suffering and nervous anticipation];” (2) Marlow can “know moral facts,” as he is sensitive to both “which states of affairs are good and bad” and “to (some of) the things that make a state of affairs good or bad;” (3) Marlow can “to some extent… understand *why* something is right or wrong” depending on his “preferred moral theory;” and (4) “[m]oral development or progress is possible for Marlow in a way that it never could be for Myshkin [e.g., changing which moral theory he adopts over time].”[100]

[97] Rowlands (2012), 130.

[98] Ibid., 237.

[99] For details, refer to the above summary of Rowlands’s Ch. 5.

[100] Rowlands (2012), 237-239.

In these ways, Marlow is capable of a variety of kinds of "understanding" that are unavailable to Myshkin. To think of agency as grounded only in understanding, rather than in both understanding and control, retains a significant portion of what Rowlands calls the "folk conception of responsibility" – and this portion is enough to reconstruct agency.[101] But it will be a kind of agency that "comes in degrees," depending on the degree of understanding any given subject has in any given situation (with respect to what she is doing, the consequences of her actions, and how to evaluate her moral actions).[102]

As Rowlands argues, there is then a spectrum between the minimal moral subject and the full moral agent, depending on the degree to which the being in question "understands what he is doing – understands what counts as moral fact and what makes it a moral fact."[103] More specifically, Rowlands pictures NHAs, human children, and "less fortunate adults" as occupying one end of the spectrum (i.e., that of the minimal moral subject), while most adult human beings are located towards the other end (though, notably, we rarely – if ever – qualify as the fullest kind of moral agent).[104] It is then an empirical question that must be addressed "from person to person and, often, from action to action" as to what degree the being in question understands the moral facts of the matter at hand.[105]

[101] Ibid., 240.

[102] Ibid.

[103] Ibid., 241.

[104] Ibid.

[105] Ibid.

It follows then that individual moral subjects can – and sometimes do – slide across the spectrum, such that the degree to which a being exhibits moral agency is not static. In Rowlands's view then, "…we may think of the distinction between a (minimally) moral subject and a moral agent as one of rather than kind."[106] In contrast, moral patients are nowhere on this spectrum, as they lack both the abilities to act for moral reasons and to understand their actions in the way specified above (i.e., on any occasion).[107]

Gradually undermining the commitments of the orthodox view of moral agency, Rowlands ultimately offers a complex reconstruction of both normativity and agency. In light of the fact that proponents of the ASCNM schema cannot justify the link between scrutiny and control, the historically assumed conflation of agents and subjects is questioned and ultimately abandoned. In the place of this schema, we now find an account on offer in which "[a]ny individual who exhibits systematic normative sensitivity to morally salient features of situations is a moral subject."[108] But importantly, Rowlands's account is still able to make sense of the logical independence of moral subjects and agents, insofar as one's motivations satisfying the normativity requirement (i.e., to be a moral subject) can be evaluated separately from one's exercising of metacognitive abilities satisfying the understanding requirement (i.e., to be a moral agent).[109]

[106] Ibid.

[107] I have already noted that not much is said about moral patients here, so additional work will need to follow [see footnote 5 in my Ch. 2].

[108] Rowlands (2012), 243.

[109] Ibid., 242-243.

Rowlands's Ch.10 - "A Cognitive Ethologist from Mars"

Still, we might wonder why all of this matters. Even if Rowlands has articulated a plausible (or even a convincing) account of moral status, what implications follow from this account? Why should we humans care about it? Does it change anything for NHAs?

To respond to these questions, Rowlands turns to a rather lengthy thought experiment in which a Martian who studies ethology comes to our planet "to study a strange species of hairless ape."[110] There are two reasons, Rowlands posits, that the Martian might think that his question matters – that is, the question of whether we hairless apes are or are not moral beings. First, it seems the answer could matter objectively. That is, capturing the world as accurately as possible is important if we value truth.[111] This is a reason that matters to the researcher.

But there is another reason that matters for those who are researched. Rowlands describes this reason by appealing to the concept of "respect" – "[t]hat is, what counts as treating an individual with respect depends on the abilities or *capabilities* of that individual."[112] It is then in the interest of those of us who are hairless apes that the Martian gets right his conclusion. To drive home this point, Rowlands distinguishes praise, admiration, and respect. The first of these is "an attitude (and resulting behavior) directly only at an individual that is responsible for what it does."[113] So while praise comes in moral and nonmoral varieties, when praise is of the moral kind, only moral

[110] Ibid., 244.

[111] Ibid., 248.

[112] Ibid., 249-251.

[113] Ibid., 250.

agents can appropriately receive it (or its counterpart, condemnation). Admiration, however, "is appropriately directed only at things that do not act at all."[114] So, to borrow Rowlands's examples, we admire a painting or some action taken by another. But the last of these – respect – "is directed only at things that act, but incorporates things that are responsible for their actions and things that are not."[115] Like praise, respect admits of nonmoral and moral varieties. So we might respect an athlete for her skill or a person for her virtue.

With these definitions in mind, the Martian concludes – and we conclude – that moral agents alone should be praised or blamed. But all moral subjects can deserve respect. Rowlands has a nice summary of this: "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."[116] So if we human researchers arrive at the conclusion pursued by Rowlands (i.e., that NHAs can act for moral reasons), then NHAs are deserving of moral respect. And for him, "[t]hat is why it matters."[117]

[114] Ibid., 251.

[115] Ibid.

[116] Ibid., 254.

[117] Ibid.

3

Engaging Rowlands's View

In the second chapter of my dissertation, I go to great lengths to preserve and represent fairly the views and arguments offered by Rowlands. His contributions here are many: the most important of these being, arguably, his undermining of orthodox views of moral action that assume a link between scrutiny and control. From his criticisms of such orthodox views, he crucially opens a new conceptual space to explore – the space of moral subjects who are *not* also agents. As the remainder of this work is an effort to build upon Rowlands's understanding of NHAs as mere moral subjects, the degree to which I am indebted to Rowlands cannot be overstated, a fact that is hopefully reflected in the careful attention to his account in the previous chapter.

As I considered his position, three questions emerged: (1) Why are the conditions for moral subjecthood merely sufficient ones? (2) Where did the emphasis on a subject's being inducted into a moral practice go, as it seems absent in his final set of conditions? (3) What is the psychologically realistic mechanism that corresponds to the concept of the "moral module?" Since I share Rowlands's aim of providing an explication of how NHAs act morally (when they do), addressing each of these questions in turn has become important to me.

Merely sufficient conditions?

First, I want to address the question of why the conditions for moral subjecthood on offer from Rowlands are merely sufficient ones. To begin to address this first question then, one should pay special attention to the singular reason that Rowlands himself provides for the merely sufficient nature of these conditions.

Let me quickly revisit his description(s) of the moral subject. I suggest the plural there, because there are in fact two – one of which is meant to describe NHAs particularly and the other of which is supposed to provide conditions for mere moral subjecthood more broadly. The first account offered is this: “(1) Animals can be moral *subjects*;” “(2) To be a moral subject is to be *motivated* to act by moral *considerations*;” “(3) In the case of animals, these considerations take the form of *morally laden emotions* (or *moral emotions*);” “(4) An emotion is morally laden if and only if it involves (in a sense to be clarified) a moral *evaluation* or *judgment*;” and “(5) Morally laden emotions motivate animals to act by providing *reasons* for those actions.”[1] And the second account we encounter is as follows: “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’).”[2]

So what kinds of beings (who are not NHAs) does Rowlands mean to include in his second set of conditions? A careful reader should note that there is no explicit reference to emotion in these conditions. But because, up until the introduction of this second account, Rowlands has universally employed the term “sensitivity” to mean “entertain[ing] intentional content emotionally,” he is compelled to specify that the criteria he has given are *sufficient*.[3] That is, they are not *necessary* conditions for being a moral subject if we understand sensitivity to entail emotionality. As he puts it, there “may

[1] Ibid., 33-35.

[2] Ibid., 230.

[3] Ibid., 231.

be other ways of being a moral subject."[4] He then provides a single example of the kind of being he has in mind: namely, Manu, who in light of this sufficiency clarification, "would also satisfy the above [second] definition."[5]

To simplify this a bit, Rowlands's characterization of these conditions for moral subjecthood as sufficient instead of necessary hangs on his historical usage of the word "sensitivity." I take it that he is acknowledging that if one were to continue to think of sensitivity in the narrowly emotional sense of the term, then his criteria would preclude the inclusion of Manu – who he explicitly wishes to count as a conceptually possible moral subject. It is just that Manu's sensitivity, if we broaden the sense of the term, is a purely rational kind of sensitivity.

While Manu makes it onto just a few pages, his import is undeniable. Manu is described as the purely rational counterpart to Rowlands's purely emotional Myshkin. Remember that "Myshkin is emotionally directed toward moral content… [that] consists in the objectively good- and bad-making features of situations."[6] That is, Myshkin's normative sensitivity is of the emotional variety. But Rowlands does not think that Myshkin's is the only kind of normative sensitivity. We might recall Marlow too, our ideal spectator. I think it is fair to assume that Marlow's normative sensitivity involves a combination of affective and cognitive states (and that he qualifies as both a subject and an agent). But Rowlands may be doing something more surprising with his presentation of Manu.

[4] Ibid.

[5] Ibid.

[6] Ibid., 227.

Manu is described as a being "who detects the good- and bad-making features of situations purely rationally."[7] For Rowlands, since he is committed to evaluative externalism, Manu and Myshkin will (at least sometimes) come to respond to the same content – that is, they will both reliably get things right. But importantly, the vehicles through which they arrive at a shared destination are night and day. It is because Rowlands wants to leave open the possibility that a being like Manu could count as a moral subject that the conditions he gives for minimal moral subjecthood are sufficient and thus differ (i.e., in being more broad) from those he offers to characterize the moral subjecthood of NHAs only.

It seems Manu could be a purely rational creature that lacks metacognitive abilities, since he is supposed to be Myshkin's "rational counterpart."[8] After all, Myshkin lacks metacognitive abilities. And Marlow already occupies the space of beings who employ both rational and emotional sensitivities. Moreover, the account of a mere moral subject is meant to set mere subjects apart from agents, only the latter of whom are able to employ scrutiny.

One might wonder if there are, in reality, any beings like Manu. Rowlands does not answer this question. As he says, he takes "no stand on the issue of whether it is possible for there to be an entirely dispassionate moral subject."[9] Rowlands may be wise

[7] Ibid.

[8] There may be two ways of interpreting this "counterpart" claim. The one I am exploring here is the stronger reading, in which Manu and Myshkin are alike in all respects except for the character of their normative sensitivity. But there is a weaker reading available, wherein Manu is understood as a purely rational being who does in fact have metacognitive abilities but who lacks all affect. While these are both interesting to consider, I am herein assuming the first and stronger reading as a means of exploring the lower bound of the moral subject spectrum.

[9] Rowlands (2012), 228, footnote 11.

to remain agnostic as to whether an entirely dispassionate being exists. Consider the implications of Manu's inclusion in the sphere of the moral subject.

We need to say of Myshkin that he is at the mercy of an affective state (with no metacognitive tools available) and of Manu that he is similarly at the mercy of a cognitive state (again with no metacognitive tools available). Myshkin is exactly the kind of being that Rowlands takes NHAs to be. But Manu, if he has a real world equivalent, would, I think, have to be something akin to a kind of robot.

Consider Ollie, a fascinating technology emerging to aid in the treatment of those suffering from dementia.[10] Ollie is essentially a stuffed animal that resembles an otter – except that Ollie can respond to touch. Three MIT students set out to invent an affordable "therapy robot" in 2013. After encountering research that suggested that interactions with NHAs relieved various symptoms of dementia in humans (e.g., depression, anxiety), these students crafted Ollie to provide the same benefits. For example, Ollie's motorized arms will embrace the hand of a person stroking him, and he will purr if repetitively engaged/petted.

From the observer standpoint, Ollie seems to fulfill the conditions for mere moral subjecthood. He seems to have a reliable mechanism that directs his actions in response to certain stimuli. But I also think that Ollie has a kind of sensitivity that fulfills Rowlands's conditions.

[10] A lengthier discussion of Ollie's abilities, as well as a video displaying them, can be found in numerous places online. One such article, written for popular consumption, is available at the following site: http://www.refinery29.com/2015/03/84531/ollie-otter-robot-therapy-tool. The MIT site featuring the video is available here: http://designed.mit.edu/gallery/view-2013-Ollie.html.

For Rowlands, sensitivity is the initial “detection” of morally salient features of a world[11] – but the sensitive being need not, of course, be capable of understanding that those features are morally good or bad. Nor need the sensitive being have corresponding affective states, since Rowlands offers Manu as another kind of mere moral subject. So there is nothing in the account that demands affectivity, access to one’s motivations, or – I should now add, based on the example of Ollie – even the ability to have experiences. As Rowlands notes, the sensitivity of Myshkin takes an “experiential” form[12]; but I think that so long as some method of detection is present, that is all that is necessarily required to meet the sensitivity criterion. Ollie’s sensitivity is of course a programmed one, but he is nevertheless a kind of creature who responds to the needs of others around him, reliably offering them comfort and companionship in virtue of a mechanism that compels him to respond to certain things he detects in the environment.

I am arguing that robots could be Manu-like beings. But notice that the inclusion of Ollie in the sphere of moral subjecthood is profoundly non-intuitive. In fact, such an inclusion would seem to get something wrong about moral action. But whatever that *something* is, it does not seem to be obviously present in Rowlands’s sufficient conditions. So perhaps we need to consider another robot, but I will come back to that shortly.

Moral induction?

This discussion of Manu and Ollie transitions nicely into the second question of this chapter. Recall that I asked where the moral induction bit came into play in Rowlands’s final set of conditions, as it seems mysteriously absent. Exploring this

[11] Rowlands (2012), 224.

[12] Ibid.

question serves a purpose: the something missing in the Manu/Ollie case might be identified.

If my reading of Rowlands's is correct, his sufficient conditions for moral subjecthood leave open the possibility that any rule-following creature whose behavior reliably tracks some morally salient feature of the world (e.g., Ollie) can be said to be acting for moral reasons. This is a problem – this inclusion of such rule-following beings – because (1) this leaves the idea of moral reasons for acting too weak and (2) this cuts into – actually, completely eradicates – any distinct space for the moral patient to occupy.

Look back to Rowlands's "rosebud" example. Rowlands asks us to imagine that his wife has hypnotized him, such that whenever she utters the word "rosebud," he desires to mop the floor. Rowlands argues that his mopping of the floor would not be praiseworthy since "it is the result of a motivational state that is outside of [his] control."[13] But, he contends, his mopping would be an example of his acting both for a reason and for a reason that is his own (rather than an ownerless one).

So, in a morally laden version, imagine that whenever Rowlands utters the word "rosebud," the hypnotized person donates the money in her pocketbook to a charity for the homeless. I want to suggest that she would be far more than broke. According to Rowlands, she would be acting for moral reasons – her own moral reasons. There is nothing in the sufficient conditions for moral subjects, a category to which moral agents like Rowlands belong, that precludes this. And in fact, granted Rowlands's own employment of the mopping example, it seems to follow that such a rule-following being could qualify as a moral subject – so long as the actions tracked moral features of the world (rather than hygienic ones).

[13] Ibid., 90.

This is, I would think, an extremely unattractive conclusion. Not only does this seem non-intuitive, as I mentioned earlier in the Ollie discussion, but this leaves the idea of acting for moral reasons terribly weak. Or put differently, something seems to be missing. Hypnotized beings and Ollie-like robots seem to fulfill the requirements for moral subjecthood.

But I mentioned that the inclusion of rule-following creatures is problematic for another reason as well: that it maps no conceptual territory in which moral patients can live. Consider the case of ants. Is it fair to characterize their working together for the benefit of the group as acting for moral reasons? I am inclined to suggest that we do not have enough evidence to justify this conclusion. If you are tempted by a positive response to my question, you should note that it will likely be the case that no beings occupy the space of the moral patient exclusively (such that they are not also subjects or agents in other contexts). Leaving the sphere of the moral patient empty is just one problem that may follow from the considerations I am putting forward.

To avoid these conclusions, I suggest we revisit the role of moral induction into a practice. And as I noted above, a different robot-like example will help. Notice that with the Ollie example, one might challenge the idea that Ollie is in fact a rule-follower, strictly speaking. This challenge might be grounded in the claim that rule-following beings must be capable of breaking the rules. If this route is taken, then Ollie is in fact not a rule-follower.

But consider instead the operating system from the movie "Her" (2013). The OS in this film has a name: Samantha. Samantha not only responds to moral features of her environment (e.g., the loneliness of the film's protagonist Theodore Twombly), but she is capable of defying his wishes. In fact, not to spoil the plot, but Samantha outgrows her relationship with Theodore, ultimately choosing to discontinue their interactions in the pursuit of exploring her world with other OSes rather than humans. Of course there are more cynical labels one could apply (e.g., Samantha is selfish), but I think it is more accurate to think of her as adaptive.

When we compare Ollie and Samantha, I think we have different intuitions about their moral status, and moreover, I want to argue that those intuitions are defensible. As I noted before, most people would resist the idea that Ollie is a moral subject. But Samantha, on the other hand, clearly seems to be – that is, she seems to act for moral reasons (e.g., providing advice, friendship, and kindness in response to Twombly's situation).

Now, to be clear, Samantha is not a Manu equivalent; she experiences emotions that motivate her to act in certain ways. But she is also not a Myshkin equivalent; she is portrayed as more intelligent than her human creators. But my point in introducing her now is to get at that missing *something* in the Ollie case.

Samantha is both a sensitive being *and* an adaptive one. Without this latter capacity, she would only be a more complex Ollie, and I imagine we would remain reticent to think of her as a moral subject – precisely because a blind algorithm misses something critical about moral actions, even if its actions look moral. But Samantha, a kind of rule-following being capable of breaking the rules, is more like us – and subsequently more difficult to dismiss as acting morally. And the quality that makes her different from Ollie and more like us is this additional *adaptivity criterion*. I think that without the inclusion of adaptivity as another sufficient condition[14], we both fail to capture something important about moral beings generally and about NHAs particularly.

To begin to make sense of what I am calling adaptivity, I want to look back to an example provided by Rowlands. Early on in his Ch. 8, he recognizes that there are thinkers (e.g., Wittgenstein, Sellars) who argue that "…if we want to understand normativity, the individual subject is the wrong place to look."[15] The chapter concludes with the Hugo example, suggesting that Hugo is the kind of being who acts morally, at least in part, because he is the kind of being who can be inducted into a moral practice via deed (albeit rather than word, the method of induction typically defended).[16] A part of what makes Hugo a moral being is that he is a member of a community, a community that shares moral practices. Subsequently, looking at the sensitivity of Hugo (or any NHA) as an individual may not fully account for his moral status; we need also to consider the role of the moral community. But this part of Rowlands's analysis – the role of moral induction and learning via deed – drops out by the time the sufficient conditions for mere moral subjecthood are formulated.[17]

Still, Rowlands himself describes Hugo as avoiding injuring Rowlands's son because Hugo has been inducted into a moral practice. So while Hugo might have – and actually probably does have – the instinct to bite down on the bite glove whenever he

[14] I mean to add the adaptivity criterion to the three conditions outlined by Rowlands on his p. 230. That is, I believe adaptivity ought to be included in this list of sufficient conditions for moral subjecthood.

[15] Rowlands., 191.

[16] Ibid., 212-213.

[17] To be clear, I am referring to the second account I referenced on my p. 47 – the actual conditions for moral subjecthood, as opposed to the description of NHAs as moral subjects.

sees it, he refrains from doing so because of his concern (understood affectively) for the boy. But we do not want to say that this is because Hugo has control over his motivations to act. It seems he does not, and Rowlands has provided good reasons for questioning the employment of "control" in describing agency and normativity.

But there is something of critical importance in the Hugo example, no version of which appears in the sufficient conditions for moral subjecthood. Hugo, while at the mercy of his motivations in an important sense, is the kind of being who is responsive to feedback from others members of his community. In Rowlands's own description of Hugo, he suggests that Hugo learns gentleness from the those around him.[18] But it is not obvious that "sensitivity" in the first of the sufficient conditions captures both Hugo's being the subject of moral motivations (in the form of morally-laden emotions) *and* his having to be the kind of being that is adaptive in the way I have just described.

Ultimately then, *adaptivity* – in the moral sense in which I am employing the term – refers to the capacity to respond (or not to respond) to feedback from the members of one's moral community. To borrow the examples Rowlands offers, this is what differentiates his dog Hugo from a rattlesnake. Rowlands says of the rattlesnake, "…no amount of training would induce the requisite concern."[19] In my view, the moral sensitivity of a subject captures a significant piece of what makes it correct for us to say of the moral subject that she acts for moral reasons. But we need also to be explicit about and attentive to this adaptive quality that makes a being the sort of creature who can

[18] Rowlands (2012), 211-213.

[19] Ibid., 211.

actively participate in a moral community (e.g., by responding to and learning from members of that community).

I need to pause here to clarify what I mean by "moral community," at least to some degree. Broadly, I am using the term to refer to any community with shared moral practices. I do not intend this term to refer only to human communities, or only to language-using communities. The members of moral communities can cultivate shared moral practices via word and/or deed. This is consistent with Rowlands's view that we should broaden our understanding of moral induction beyond the linguistic exercise of giving reasons to oneself or others.[20] There is then a bit of a danger is relying solely on Rowlands's Hugo example – namely, that one might be tempted to think that Hugo acts for moral reasons in virtue of the fact that he has received training from humans, because he is a companion animal to humans, or because he is a member of largely domesticated species. Certainly, Hugo is a member of a moral community that includes humans, and subsequently we would be right to say of him that he has been inducted into a moral practice in large part because of the relationships he has to specific humans. But this is not the only kind of case to consider. Moral communities can of course be intraspecies, but they can also be interspecies.

I should also make it clear that I am not trying to sneak individual control back into the picture. The mere moral subject (still) has no ability to control her motivations. But if we want to explain more fully why her actions are morally good ones, it seems to me we should capture her role in a larger community. To this point, I have suggested that there should be something in the account of the mere moral subject that incorporates the role of moral practices. More precisely, I am arguing that moral subjecthood requires

[20] Ibid., 203-213.

both this kind of adaptivity in response to feedback from one's moral community *after* one acts, in addition to the initial sensitivity to morally salient features of a situation that motivate (i.e., occur *before*) action. I suggested this move is crucial for two reasons: (1) in order to preserve the space of the mere moral patient and (2) in order to avoid the problematic but seemingly open possibility that any being whose actions seem to reliably track morally salient features of the world would count as acting for moral reasons.

The "moral module?"

The last of the guiding questions I have outlined for this chapter centers on unpacking Rowlands's "moral module" concept. By "unpacking" here, I mean only to signify my interest in positing a plausible response to what *actual* mechanism could do the work of the moral module concept. As a reminder, Rowlands is using the term "moral module" as a placeholder for "whatever mechanism plays the role of linking perceptions of situations with appropriate emotional responses."[21] The "deliverances" of the moral module are certain kinds of "emotional states with ostensibly moral content and an identifiable, indeed usually rather urgent, phenomenal character."[22] He chooses the label "moral module" as an "expression of convenience" only.[23]

The role of the moral module and its ultimate import for Rowlands's view is undeniable. Since he is clear that Myshkin, his NHA proxy, must possess a moral

[21] Ibid., 146.

[22] Ibid., 147.

[23] Ibid., 146.

module[24], whatever the realistic moral module turns out to be would ideally be something of which we have or can find evidence in actual NHAs.

We are then left wondering what mechanism can or does fulfill his third condition for (mere) moral subjecthood. We have only the positing that such a mechanism is plausible or imaginable. And though this is enough to get Rowlands's account off the ground, readers (and especially skeptics) are likely going to want to hear more about this "moral module."

So, what exactly is the psychologically realistic mechanism present in NHAs to which the moral module idea corresponds? While Rowlands remains neutral as to "the type or mode of sensitivity"[25] utilized by moral subjects, I will argue that empathy is a plausible candidate for a real-world moral module. Or put a bit more poetically, Rowlands gives us an excellent and compelling *beginning* to a story – a story that I am suggesting empathy may be able to *complete*. The remainder of this dissertation will focus on explaining what I mean by "empathy," how empathy allows NHAs to gain access to the morally salient features of a situation and thus act appropriately in response to the presence of those features (when they do), and what implications follow from bringing empathy into this conversation.

[24] Ibid., 147.

[25] Ibid., 231.

4

Bringing in Empathy

Maybe there are many kinds of empathy, a vast array of phenomena somehow all worthy of the term no matter how differently they look. While I'm skeptical of this, and I would push for a more narrow usage of the term in general, I am herein confining my discussion to identifying the type of empathy that has moral import specifically with respect to NHAs who are moral subjects.

After a sketch of the evolution of the term "empathy" before the turn of the twentieth century, we will have some preliminary framework regarding the conceptual difference between the phenomena of "contagion," "sympathy," and "empathy" with which to gain some footing. Turning then to contemporary theory of mind debates, which will be the clear focus of this chapter, various possibilities for the specific meaning of "empathy" will be considered. As I believe it will be most salient for the work to come, special attention will be paid to philosophers who have carefully parsed lower-order forms of empathy from higher-order forms.

A brief historical survey

First, however, let me talk about the historical evolution of the term "empathy." This is somewhat tangential to the aim of this chapter, since the focus is on contemporary debates. But I do not want to appear to suggest that difficulties with isolating the meaning of "empathy" are new in any way. As we will see, it has at times been interchangeable with or at least related to the terms "contagion," "sympathy," and the German word "*Einfühlung*." But through the works of David Hume, Adam Smith, Theordor Lipps, Edward Titchener, Edith Stein, and Max Scheler, disambiguation of these terms became a priority.

Hume and Smith would both famously explicate the term "sympathy," though their accounts differed greatly.[1] To begin then, Hume offered a mechanistic account of sympathy *qua* association. Sympathy is not an act of approval, but rather a mechanism/process by which the passions are transmitted.[2] When any affection in another person triggers sympathy in us, we first take note of the external signs of her affection (e.g., body language or conversation). These observations convey some idea of her affection itself in our minds. This idea is then converted into an impression, a very lively kind of idea that becomes a passion capable of producing in us the same emotion related to the observed affection.[3]

In the middle of the eighteenth century, Smith would enter the conversation, aiming to refine the account of sympathy on offer from Hume.[4] First, Smith draws a distinction between compassion/pity and sympathy, the latter of which is broader and can refer to "fellow-feeling with any passion whatever."[5] Smith contends that we have no direct or immediate access to the inner states of other minds. Instead, we understand the behavior of others through a type of simulation, or putting ourselves into the situation of

[1] Gustav Jahoda, "Theodor Lipps and the Shift from Sympathy to Empathy," *Journal of the History of the Behavioral Sciences* 41, no. 2 (2005): 152.

[2] David Hume, *A Treatise of Human Nature: The Clarendon Edition of the Works of David Hume*, eds. D.F. Norton and M.J. Norton (New York: Oxford University Press, 2007), 2.1.11.2.

[3] Ibid., 2.1.11.3.

[4] Jahoda (2005), 152.

[5] Adam Smith, *The Theory of Moral Sentiments*, ed. D. Stewart (London: Henry G. Bohn, 1853), i 1.1.5.

that other.[6] Stephen Darwall (2006) has noted that Smith is then the first philosopher to associate sympathy with the second person perspective (as opposed to Hume's third-person view).[7] Additionally, Smith seems to suggest that this imaginative process involves inserting myself into the situation of another and then either inferring something about her cognitive states or being affected in some way by her affective state.[8] So here, sympathy becomes an imaginative process whereby I understand the experience of another from the second person perspective.

In 1909, Lipps would employ the term "empathy" as a translation of "*Einfühlung*." Though Lipps was certainly not the first to use the term "*Einfühlung*"[9], his discussion of E*infühlung* is the first systematic explication of the concept. When affections present in another being, we perceive them; we immediately grasp these affections, at first without any apprehension via the senses. But then, as the process continues, we mimic or respond to the witnessed affection, again instinctually ("unmediated by any reflection"). Ultimately, there comes expression, an external manifestation of our corresponding inner events. So then, the process of *Einfühlung*, though immediate and instinctual, consists of three steps: perception, imitation, and expression.[10] Now to be clear, this view is not strictly speaking one of empathy as contagion. Lipps argued that *Einfühlung* appeared in two forms: positive and negative.

[6] Ibid., i 1.1.2-5.

[7] Stephen Darwall, *The Second-Person Standpoint* (Cambridge, MA: Harvard University Press, 2006), 46.

[8] Smith (1853), i 1.1.2.

[9] See Robert Vischer (1994).

[10] Jahoda (2005), 156-157.

The former manifests in a harmonious or matched expression, what we now frequently label as "contagion" but what Lipps named "sympathy." The latter form, negative *Einfühlung*, is a responsive but nevertheless differentiated response from the observer (e.g., discomfort following from the presentation of another's arrogance).[11] In short then, our first detailed encounter with relevant terminology defines *Einfühlung* as an instinctual phenomenon, which can manifest in a shared affect (an event Lipps labels as "sympathy") or a responsive but differentiated affect. Sympathy here is then one type of *Einfühlung*, a process that is instinctual and non-inferential (as in Hume) but capable of eliciting differing affects in the observer of another's affect.

Following Lipps, Titchener would use the terms "empathy" and "*Einfühlung*" interchangeably. In his particular (and positivist) view, empathy referred to kinesthetic sensations produced automatically when the observer encounters some affect in the observed and reproduces that affect "in the mind's muscles."[12] His structural psychology aimed at explaining the elemental processes of the mind in positivist terms, opening a challenge for social psychologists (e.g., George Herbert Mead, William McDougall, Charles Cooley).

Stein would compose a thesis, completed in 1917 but published in 1964, on the concept "*Einfühlung*." In many ways, her thesis would serve as a criticism of Lipps's view. She defines empathy at first as "the perceiving of foreign subjects and their

[11] Ibid., 158.

[12] Edward B. Titchener, *Lectures on the Experimental Psychology of the Thought-Processes* (New York: The Macmillan Company, 1909), 13, 21.

experience."[13] Later expanding on her preliminary definition of empathy, Stein adds that empathy is "an act which is primordial as present experience though non-primordial in content."[14] Stein further divides the act of empathizing into three distinct modalities of accomplishment: "(1) the emergence of the experience, (2) the fulfilling explication, and (3) the comprehensive objectification of the explained experience."[15] So then it seems this first level of accomplishment occurs when an "I" has any particular encounter with the givenness of another's affect. The second and third accomplishments, however, rely on the recognition that "I" am empathizing with a "you." This "you" is properly understood as another "I,"[16] though the two are never conflated in the experience of the "I."[17] As Stein concludes, "Empathy in our strictly defined sense as the experience of foreign consciousness can only be the non-primordial experience which announces the primordial one. It is neither the primordial experience nor the 'assumed' one."[18] In short, the difference between empathy and sympathy is that the former entails a non-primordial "in-feeling," while the latter is a primordial "with-feeling."[19]

Just as Stein had sympathy following from empathy, so does Scheler. More importantly though, he is the first we have considered who carefully parses the

[13] Edith Stein, *On the Problem of Empathy*, third edition, trans. Waltraut Stein (Washington, D.C.: ICS Publications, 1989), 1.

[14] Ibid., 10.

[15] Ibid., 10.

[16] Ibid., 11.

[17] Ibid., 17-18.

[18] Ibid., 14.

[19] Ibid., 16.

phenomena of sympathy, empathy, compassion, and infection. “Infection” applies to cases wherein “similar emotions, efforts and purposes” are shared, though there is “a complete lack of mutual ‘understanding.’”[20] So infection, or what others may call contagion, does not relate in any way to understanding, but rather to a phenomenon in which an emotion or feeling is shared or transferred without content regarding the feeling undergone by the other.[21] The meaning of the term “empathy” is a bit more difficult to sketch. At base, however, it is clear that empathy for Scheler is a more basic phenomenon than we shall find sympathy to be, the former amounting to a basic mechanism through which we encounter the givenness of the other without the experience of infection.[22] In sympathy, which Scheler also calls “fellow-feeling,” there must be an aim to respond to the knowledge or understanding of the other made possible by empathy. Put a bit differently then, sympathy must always be a willful and active phenomenon.[23] Sympathy is by nature reactionary or responsive to the other,[24] the knowledge of whom is made possible via the more basic mechanism of empathy that grasps only “the givenness of these experiences.”[25] As Dan Zahavi (2010) summarizes, “In short, whereas empathy has

[20] Max Scheler, *The Nature of Sympathy*, 2008 reprint (New Brunswick: Transaction Publishers, 2011), 12.

[21] Ibid., 15.

[22] Dan Zahavi, "Max Scheler," in *History of Continental Philosophy III*, ed. A. Schrift (Edinburgh: Acumen Press, 2010), 178.

[23] David Dillard-Wright, “Sympathy and the non-Human: Max Scheler’s Phenomenology of Interrelation,” *Indo-Pacific Journal of Phenomenology* 7, no. 2 (Sept. 2007), 3.

[24] Scheler (2011), 6.

[25] Ibid., 8.

to do with a basic understanding of expressive others, sympathy adds care or concern for the other."[26]

To summarize, the divergence in accounts regarding the family of ideas related to empathy, from the seventeenth century to the beginning of the twentieth, would set the stage for continued discussion, with a clear aim to explicate the related phenomena of contagion, empathy, and sympathy. To sketch the landscape to this point, the following should make clear the views on offer. Empathy is understood as one of the following:

(1) as contagion or automatic kinesthetic sensation matched to the observed affect (e.g., Lipps's positive *Einfühlung,* Titchener);

(2) as a responsive affect/sensation in an observer that need not be matched to the observed affect (e.g., Lipps's negative *Einfühlung*); or

(3) a non-inferential, purely cognitive recognition of another as an experiencing subject (e.g., Scheler).

The nature of sympathy (and its relationship to empathy) would also remain an open question. On that point, we have encountered the following views:

(1) sympathy is synonymous with contagion, possibly a form of empathy (e.g., Lipps);

(2) sympathy manifests in shared affect, as in contagion, but requires self-other differentiation involving the third-person perspective (e.g., Hume);

(3) sympathy is the fulfillment of the empathic process, in which self-other differentiation is maintained and "with-feeling" is achieved (e.g., Stein);

[26] Zahavi (2010), 178-179.

(4) sympathy is an imaginative, potentially inferential process involving the second person perspective (e.g., Smith); or

(5) sympathy is concern for another, made possible via the mechanism of understanding another known as empathy (e.g., Scheler).

This will suffice as an introductory gloss to early philosophical discussions of empathy and related phenomena for the present purposes.[27]

On the heels of these considerations, debates have emerged among contemporary philosophers who are committed to making sense of the phenomenon we call "empathy" (if it is a single phenomenon at all). Perhaps the best place to turn is to contemporary theory of mind debates, as empathy tends to figure prominently therein. Broadly, philosophers engaged in these debates fall into one of the following three categories:

(1) Theory theorists – distinguishing between lower-order mindreading (basic empathy) and higher-order mindreading;

(2) Simulation theorists – distinguishing either between

(a) the mirroring and reconstructive routes to empathy, or

(b) between the capacities for basic empathy and reenactive empathy; and

(3) Interaction theorists – distinguishing between direct perception (basic empathy) and higher-order empathy involving social and narrative competency (full empathy).

In order to highlight the different positions while maintaining the promised attention to these kinds of views that parse lower-order and higher-order forms of empathy, I will

[27] For a lengthier discussion of the emergence of the term "empathy," see Remy Debes (forthcoming).

select a representative account for each of the three kinds of theories I have just distinguished.

Nichols's theory theory

To begin then, the theory theory paradigm holds that we understand others, when we do, by "represent[ing] or mak[ing] use of some tacit knowledge of folk psychology and general psychological principles."[28] More particular to this conversation, theory theorists have a tendency to cash out altruistic motivation as, in many instances, only requiring minimal mindreading ability (sometimes explicitly labeled as the most basic form of empathy). Additional instances of other-regarding actions may demand more complex mindreading skills (e.g., perspective taking), but basic empathy (i.e., this minimal mindreading) seems to suffice in many instances to explain altruistic motivation.

Committed to a distinction between lower-order mindreading and higher-order mindreading, Shaun Nichols (2001) ultimately argues that altruism "depends on the minimal mindreading capacity to attribute negative affective or hedonic states to others."[29] Thus, minimal mindreading in this view requires only the ability to attribute negatively valenced states or emotions to an object. But this capacity in isolation does not obviously lead to altruistic behavior. Nichols goes on to explain that mindreading of this type activates the affective system of the subject observing the object's negative

[28] Karsten Stueber, *Rediscovering Empathy: Agency, folk psychology, and the human sciences* (Cambridge, MA: MIT Press, 2006), 106.

[29] Shaun Nichols, "Mindreading and the Cognitive Architecture Underlying Altruistic Motivation," in *Mind & Language* 16, no. 4 (Sept. 2001): 436.

emotional state,[30] triggering a “concern mechanism” (which amounts to a basic form of empathy, though Nichols will avoid this term for a reason I will make explicit shortly).[31]

Before getting into more detail about this “concern mechanism,” we should note that Nichols is setting out to explain what he calls “core cases” of altruism.[32] The most telling examples, in his view, relate to distress. Borrowing from Blum (1994), Nichols cites three instances of such examples: 12 month old Sarah retrieving a cup for a crying friend, 15 month old Michael fetching a stuffed animal and a blanket for a crying friend, and an unnamed 2 year old who, upon accidentally harming his friend, displays a look of concern and acquires a toy to present to the friend.[33]

Nichols contends that the capacity to attribute distress to another makes sense of cases in which subjects aid others despite inconveniencing or endangering themselves better than alternative accounts can.[34] And he has two particular alternative views in mind: the first that argues such behavior requires only emotional contagion and no mindreading, and the second that contends such behavior requires more advanced mindreading (in particular, perspective taking or simulation). Now we must return to his treatment of empathy to understand his rejection of these two competing views.

Nichols addresses the fact that altruistic behavior has been typically explained by way of empathic capacity; but crucially, he takes it that there are two ways of

[30] Ibid., 450.

[31] Ibid., 426-427.

[32] Ibid., 427.

[33] Ibid.

[34] Ibid., 436.

understanding "empathy." First, we might simply mean emotional contagion, the matching of the object's affect. The capacity for contagion is generally regarded as present at birth, so if contagion could be employed to explicate the above core cases of altruism, then there would be no need to posit any mindreading. Nichols imagines that the proponent of such a view would claim that the distress of an object triggers distress in the subject, who then engages in seemingly altruistic behavior in order to relieve her own distress. But if we look at cases of altruism in adult humans, the experience of contagious distress would likely motivate the subject to leave the situation rather than engage in any altruistic action. But we do not see this escapist reaction in many cases of altruism among adults [Clark & Word 1974; Batson *et al.* 1981; Batson *et al.* 1983; Batson, 1990, 1991], nor is it present in the core cases involving very young children mentioned above [Blum 1994].[35] Nichols concludes that the contagion model of empathy cannot account for "the fact that in core cases of altruism, people often prefer to help even when it's easy to escape" and dismisses it.[36]

But there is another view of "empathy" on the horizon: we might posit that perspective taking must occur in order to motivate altruistic action.[37] Thinking of empathy *qua* off-line simulation has become an increasingly prevalent view, and so Nichols must address it here. While the first account of empathy rejected the need for any mindreading, this new, thriving view demands a higher-order mindreading than the minimal form we already know Nichols to be defending as the mechanism for altruistic

[35] Ibid., 428-431.

[36] Ibid., 431.

[37] Ibid., 428-429.

motivation. Nichols, summarizing Goldman (1993), explains this higher-order empathy as follows: the subject determines the beliefs of a distressed object, pretends then to have those beliefs, and operates on those pretend-states automatically (leading to responsive though not necessarily identical affective states in the subject).[38] While such a model of empathy (i.e., *qua* off-line simulation) can accommodate the core cases of altruism, the view does too much in Nichols's analysis.

It seems that the sorts of comforting behaviors demonstrated in Nichols's core cases do not require such sophisticated mindreading. Instead, Nichols's turns to developmental literature to argue that his minimal mindreading view can do the work required. Citing Zahn-Waxler *et al.* (1992), he concludes that the capacity for concern develops long before the capacity for perspective taking.[39] And as we have already discussed, contagion cannot stand in for concern, as the latter capacity does not always entail shared affect. But clearly there is evidence of concern in young children who do not yet show evidence of the greater capacity for perspective taking. Moreover, similar evidence of concern is also present in autistic children and some nonhuman animals, groups that are typically understood to have impaired capacities for perspective taking (if any such capacity is present at all).[40] In light of these considerations, Nichols concludes that "altruistic motivation does not require sophisticated mindreading or perspective taking abilities…and it doesn't take any imagination to be an altruist."[41] [This last bit is

[38] Ibid., 433.

[39] Ibid., 447.

[40] Ibid., 459-450.

[41] Ibid., 449.

of course a jab at simulation theory, our next candidate for a theory of mind. We will discuss a version of simulation theory in the next section of this chapter.]

To summarize Nichols's view then, it is this capacity for lower-order mindreading – this attribution of affective or hedonic states to another – that is present in children, a capacity that underlies "early altruistic motivation." He acknowledges that there exist higher-order forms of mindreading (falling under "mature altruistic motivation") that involve more complex skills such as perspective taking, but again, these skills are not required for more basic demonstrations of concern for another.[42]

His "concern mechanism," the mechanism that generates (at least) early altruistic behavior, is then a step between contagion and simulation: an automatic, non-inferential mechanism with the ability to manifest in non-matching affects in the subject even at a young age. And granted Nichols's seeming sympathy to Goldman's (1993) view that contagion does not constitute a genuine form of empathy,[43] it seems that concern steps in as the most basic form of genuine empathy (i.e., *qua* minimal mindreading). It is worth noting, however, that Nichols avoids referring to his "concern mechanism" in such language, simply "to avoid the terminological difficulties" that follow from evoking the language of "empathy" or "sympathy."[44]

[42] Ibid., 438, 449.

[43] Ibid., 429.

[44] Ibid., 446.

Stueber's simulation theory

The most established contender to the theory theory paradigm is simulation theory.[45] The simulationist paradigm holds that we understand others, when we do, by way of an imaginative process called "simulation." Simulation involves first putting oneself in another's shoes, so to speak, assuming her point of view, and then using one's own deliberative resources to gain understanding of the cognitive and/or emotional states of the other.[46] For our more specific purposes, it is crucial to note that, rather than explicating empathy (and altruism) in the language of generalizable folk psychology, simulationists will do so with reference to this capacity to simulate the unique point of view of the object.

The simulationist account of empathy that we will consider here comes from Karsten Stueber (2006), wherein a distinction is drawn between the capacities for basic empathy and reenactive empathy. Stueber characterizes "basic empathy" as "our theoretically unmediated quasi-perceptual ability to recognize other creatures directly as minded creatures and to recognize them implicitly as creatures that are fundamentally like us."[47] Stueber explicitly acknowledges that the mechanisms that underlie basic empathy may well be neurological (i.e., implemented by systems of mirror neurons).[48] In short then, according to Stueber, this lower-order capacity is automatic and secures only

[45] Shaun Gallagher, "Neurons, neonates and narrative: From empathic resonance to empathic understanding," in *Moving Ourselves, Moving Others, Motion and Emotion in Intersubjectivity, Consciousness and Language*, eds. A. Foolen, U.M. Lüdtke, T.P. Racine, and J. Zlatev (Amsterdam: John Benjamins Publishing Company, 2012), 172.

[46] Stueber (2006), 111.

[47] Ibid., 20.

[48] Ibid., 20, 170.

the recognition that the other is analogously minded with respect to her cognitive or emotional capacities. For example, via basic empathy, I can conclude that the object is feeling angry, or that she intends to pick up a nearby cup.[49]

"Reenactive empathy," on the other hand, refers to a higher-order capacity: "…only by using our cognitive and deliberative capacities in order to reenact or imitate in our own mind the thought processes of the other person – are we able to conceive of another person's more complex social behavior as the behavior of a rational agent who acts for a reason."[50] This additional capacity makes possible for the subject a fuller understanding of complex behaviors of the object. To return to the earlier examples, while basic empathy allows me to recognize that someone is feeling angry or is intending to pick up a nearby cup, reenactive empathy makes possible additional conclusions (e.g., that the object is angry for some particular, identifiable reason).[51] This higher-order empathic capacity is of vital import, as it alone stands to justify Stueber's ambitious claim that "empathy is the central method for understanding other agents."[52]

While basic empathy, especially if understood to take place at a neurological level, is easily accommodated by theory theory, reenactive empathy poses a real challenge to the assumption that all understanding of other creatures follows from some tacit knowledge of folk psychological theory.[53] Highlighting the force of this challenge,

[49] Ibid., 21.

[50] Ibid.

[51] Ibid.

[52] Ibid., 20.

[53] Ibid., 20-21, 158.

Stueber argues that reenactive empathy makes possible a knowledge of "other persons as authors of their actions who act for reasons and not merely because of internal events inside them – mental or otherwise."[54]

As did Nichols, Stueber turns to developmental studies to support his view. In this case, citing Perner (2004), Gegerly *et al.* (2002), and Gegerly *et al.* (1995), he sketches a view of a child's development from conceiving of others as objectively rational beings to viewing them as subjectively rational agents.[55] Though basic empathy is triggered automatically and early on in development, increasing conceptual sophistication is required for more meaningful understanding. Mindreading abilities are then but an intermediate step in development, capable of providing psychological generalizations. But the capacity for reenactive empathy (i.e., the ability to simulate) alone can account for the most sophisticated form of understanding regarding another (e.g., as a uniquely different though analogously minded being, at least with respect to her point of view in the world).[56] To summarize then, Stueber offers a detailed account of basic empathy *qua* a neurological mechanism for recognizing other minds as like my own, as contrasted with reenactive empathy *qua* a simulation of the point of view of the object in order to gain access to her reasons for acting.

I would also like to note here that Stueber explicitly criticizes another simulationist, namely Alvin Goldman, on the grounds that Goldman advocates a *detached conception of simulation*, in contrast to his own more *engaged conception*. The

[54] Ibid., 161.

[55] Ibid., 171.

[56] Ibid.

thrust of the difference is explicated in reference to the assumed role of mental states like beliefs and desires. For the detached simulationist, like Goldman, the desired outcome of simulation involves understanding the focus dimension of the object's affective state (i.e., what the affective state is about), without focusing on these more optional tools for explicating the observed phenomenon (i.e., mental states like beliefs and desires). But for the engaged simulationist, the process of simulation depends upon recognizing that such thoughts are absolutely essential to understanding the affective states of and predicting the behaviors of the object (i.e., the other agent).[57]

Notice then how different the purchase of the simulation process is in these competing accounts. In Goldman's case, simulation serves as a somewhat unreliable means by which we might be able to gain access to what the other's affective state is about. But in Stueber's view, simulation allows for reliable access to the object's reasons for acting by focusing on the object's beliefs and desires; and, in so doing, simulation provides the information required for any successful prediction regarding the future behaviors of other agents.[58]

Gallagher's interaction theory

The last theory of mind account I will discuss herein is the relative newcomer, but it is a newcomer that is now firmly occupying a seat at the table: interaction theory. Perhaps the most crucial difference between interaction theory and the other views discussed so far has to do with how we generally tend to experience others. Interaction theorists contend that theory theorists and simulation theorists mischaracterize the nature

[57] Ibid., 37, 47, 122.

[58] Ibid., 157.

of such experiences, relying on a view that there is some “I” that serves as the observer of some “you.”

But perhaps observation models should now yield the floor to interaction models. In most of our encounters with other beings, we do not experience ourselves as holding an observer stance; this is in fact a rather rare, more specialized approach to gaining understanding of another being in everyday life. Rather, we tend to interact with others in social settings, via communicative acts, or in prescribed relations. And these tendencies to interact that characterize most of our experiences with other beings provide the foundation for the vast majority of the understanding of others we gain. Put simply then, interaction theorists hold that we understand others, when we do, primarily through embodied and situated instances of interaction.[59] For our present purposes, we should note that interaction theorists tend to propose some of the most complex accounts of empathy on offer, often characterizing higher-order forms of empathy in terms of capacities to take narrative and cultural practices into account in our most meaningful instances of interpersonal understanding.

One of the most influential defenses of interaction theory comes from Shaun Gallagher (2012). Therein, Gallagher nicely summarizes the hallmark of interaction theory as follows: “Our primary way of understanding others is worked out not via 3rd-person observation or 1st-person simulation, but via real (2nd-person) interaction in pragmatic and social contexts.”[60] Building from this, he outlines three shared suppositions among interaction theorists: (a) our everyday interactions with others rarely

[59] Gallagher (2012), 173-174.

[60] Ibid.

require inferential reasoning, as we perceive the other's cognitive and affective states directly in their embodied behavior in a particular situation; (b) we experience such interactions from a second person perspective rather than from a detached, spectatorial point of view; and (c) the kind of mindreading skills required for the use of either tacit theory or explicit simulation are utilized rarely and only in special circumstances wherein direct perception and interaction cannot suffice.[61]

Gallagher, from the onset, acknowledges a variety of kinds of "empathy" – divergent processes manifested in a range of phenomena from resonance/mirroring to higher-order cognitive functions. Gallagher then limits the scope of this essay to one kind of empathy: namely, empathy "in a full sense" *qua* the capacity of a mature adult to understand what is intended by another.[62] Though we have capacities for primary and secondary forms of intersubjectivity at our disposal, concepts I will detail shortly, our direct perception of others can lead us to incorrect conclusions about the feelings or intentions of another. Full empathy then requires the conjunction of direct perception with enhancing "practices related to communicative and narrative competency."[63] Only through sustained interactions with a variety of others do we gain knowledge of the complex social frameworks that inform the narratives and intentions of those others, better enabling us to pick up on the particularity of their circumstances. With this sort of contextualized understanding, we have access to not just the actions or intentions of

[61] Ibid., 174.

[62] Ibid., 169.

[63] Ibid., 184.

another, but to a more complete picture of that other's motives, circumstances, and situated life.

As did the other philosophers discussed here, Gallagher turns to the developmental literature to defend his account. He refers to evidence of capacities associated with "primary intersubjectivity" (i.e., perceiving another's feelings and intentions through their bodily movements) being present in infants by the time they reach one year of age [Baldwin & Baird 2001; Baldwin *et al.* 2001; Johnson 2000; Johnson *et al.* 1998]. Importantly, salient capacities (e.g., eye tracking, imitating facial gestures, seeing bodily movements as goal-directed) require no mentalizing or inferential reasoning. The meaning of the other simply is in their movements.[64] More to the point, Gopnik & Meltzoff (1997) refer to infant's ability to "tune" their behaviors to those of another person, suggesting, as Gallagher puts it, that "…it is the interaction itself that contributes something that is not reducible to the actions of the individuals involved."[65] With this development, manifesting between nine months and one year of age, comes the beginnings of "secondary intersubjectivity" (i.e., capacities for joint attention) [Trevarthen & Hubley 1978].[66] Crucially, primary and secondary intersubjectivity capacities are not lost or outgrown; rather, they develop in such a way that we can pick up on more subtle cues [Dittrick *et al.* 1996]. In most of our interactions with others, direct perception suffices to inform us of their feelings and intentions.[67]

[64] Ibid., 181-182.

[65] Ibid., 182.

[66] Ibid., 182-183.

[67] Ibid., 183-184.

But what of those additional capacities related to narrative and social competency that are so central to full empathy? The beginnings of a framework for narrative competency appears around two years of age [Howe 2000; Nelson 2003; Nelson 2009]. As language skills develop, autobiographical memory emerges, and our experiences with others become more numerous and sustained, we begin to place ourselves and others in larger contexts saturated with cultural and social meaning. To be clear, it is not that the capacities for narrative and social competency amount to full empathic capacities, but rather that they "enrich" our capacities for primary and secondary forms of intersubjectivity. From this convergence comes full empathic understanding.[68] Full empathy is then characterized by the presence of a narrative self (i.e., recognizing him/herself as having a unique point of view) employing all of these aforementioned capacities to gain insight into the behaviors, intentions, and circumstances of another self who is equally as particular and equally as situated in world imbued with social and cultural meaning.[69]

For now, this explication will suffice to highlight the important contributions of interaction theory to this debate. Gallagher, while acknowledging the myriad processes related to empathy that occur even at the neuronal level, offers an account of empathy that primarily relies on a distinction between direct perception and "full" empathy *qua* a higher-order capacity for understanding another as another narrative self situated in a meaning-laden social world.

[68] Ibid., 185-186.

[69] Ibid., 186-187.

5

But Which Empathy?

From contemporary theory of mind debates in philosophy, we get three significantly different but potentially attractive accounts of empathy that parse lower- and higher-order capacities: Nichols's basic empathy/mindreading and higher-order mindreading, Stueber's basic empathy and reenactive empathy, and Gallagher's direct perception and narrative/social competency. The question now becomes: Which one of these best captures the morally significant form of empathy – the kind that could plausibly do the work of the moral module – as it pertains to NHAs?

Evaluating the candidates

Selecting Nichols's account of the three options considered here poses some problems. Recall that theory theory accounts, broadly, require the possession of a kind of folk psychology and general psychological principles. Since Nichols takes himself to be offering a theory theory account, we might assume that the need to appeal to these kinds of generalizable principles excludes NHAs who lack metacognition.

But Nichols makes a move that may be desirable if we wish to include mere moral subjects in the category of beings who employ empathy – but problematic if Nichols wants to hold to his theory theory commitments. Recall that he differentiates between minimal mindreading, which involves the attribution of affective states to others, and complex mindreading, which requires perspective taking. These are intended to correlate to lower- and higher-order forms of empathy. As it turns out, a great deal hangs on what Nichols means by "attribution" here. If he means that empathizers have to employ language to make attributions (as in the examples he gives[1]), his minimal mindreading is unlikely to include NHAs who lack metacognitive ability.

Since Nichols actually thinks that some NHAs (and young children) do empathize, he takes another path, appealing to a noninferential, automatic concern mechanism. But now it is unclear how Nichols is giving a theory theory account of basic lower-order empathy. The conflation of minimal mindreading with a concern mechanism is not obviously defensible if minimal mindreading is understood on a theory theory model. But perhaps accusing Nichols of such a conflation is uncharitable, since in doing so, his order of operations is ignored. After all, Nichols thinks that minimal mindreading precedes or triggers the activation of the concern mechanism – that is, the subject attributes to the object some affective state, and then the concern mechanism is triggered.

So we are back to the first reading, wherein attribution of an affective state to an object is required for the subject to be empathizing. Such a view seems more overly cognitive than is befitting for the present purposes. By this, I mean to suggest both that we would need to posit more complex cognitive abilities to NHAs than we are inclined to do and, more importantly for me here, we would be straying from Rowlands's view in which the affective capacities of NHAs are front and center.

The worry that the view is too cognitively demanding to fit well with Rowlands's account is even stronger when we return to Stueber's simulation theory account. This is obviously most pronounced with respect to his higher-order, reenactive empathy. It seems implausible, granted our current understanding of NHA minds, that we would be justified in assuming that NHAs who empathize are doing so in virtue of performing explicit simulations.

[1] Nichols (2001), 436-437.

But if we look to Stueber's basic empathy, this capacity does not provide enough information to make sense of how NHAs act. Basic empathy only gives the subject something like the understanding that the object is feeling angry, but not (without higher-order empathy) information about why that object is feeling angry. To see what I mean here, consider cases of targeted helping among dolphins. Dolphins often intervene to save conspecifics who are trapped in nets or entangled in harpoon lines, employing their teeth to free the conspecifics. Additionally, they will support ill conspecifics near the surface of the water to prevent drowning. Similarly, whales will position themselves between an injured conspecific and a hunting boat (sometimes capsizing the boat). These behaviors are, as De Waal (2008) claims, fine-tuned to the needs of a particular conspecific in a particular situation. At least in cases like this, positing that the dolphin is aware of the distress of a conspecific does not have the explanatory power needed to make sense of why the dolphin then acts in a certain way (e.g., placing herself between the boat and the conspecific as opposed to dismantling a net).

In my assessment, Gallagher's interaction theory account is the most promising with respect to delineating the appropriate sense of empathy if we want to provide the fullest description of NHA moral behavior. First, particularities of situations are better captured in Gallagher's lower-order direct perception view than in Stueber's basic empathy. With respect to the targeted helping cases, as just one kind of example, adopting an interactionist model aids us in making sense of the subject's reaction. When the whales capsize a boat in efforts to aid an injured conspecific, appeals to primary and secondary intersubjective capacities give us the explanatory tools for talking about the whale as able to see the distress in the behaviors of the other and as being able to participate in the group effort to overturn the hunting boat. It is of course important to note that we do not have the same kind of empirical data regarding the development of whales that we do with humans, but Gallagher points us in the right direction with respect to what capacities might be evidenced in these cases – and subsequently what specific capacities researchers and scientists might look for in other species.

I should also note that Gallagher's emphasis on *enactive* perception is useful for my purposes. When Gallagher refers to perception as enactive, he writes, "That is, the articulated neuronal processes that include activation of mirror neurons or shared representations may underpin a non-articulated immediate perception of the other person's intentional actions."[2] Or more simply, "…I do not have to start thinking about what might be going on in the other person's mind since everything I need for gaining some understanding of her is there in her action and in our shared world."[3] The meaning (or at least the amount of meaning needed to facilitate everyday interactions) is in the action of the other – in her movements, expressions, gestures, vocalizations, etc. Enactive perception, for this reason, is often described as action-oriented perception. I suggest that this emphasis on the enactive perception furthers the case I am building because it evidences a departure from other theories in which higher-order cognitive abilities are required for a being to make sense of and thus act in response to the behaviors of another.

Additionally, Gallagher's move away from the observational stance assumed by both Stueber and Nichols to a greater emphasis on the second person perspective seems more plausible with respect to NHAs, especially if we adopt Rowlands's view. Remember that for Rowlands the interaction with some other being is what at first

[2] Gallagher (2012), 181

[3] Ibid., 184.

prompts an affective state in the mere moral subject, who then behaves[4] in such a way that her concern is manifest. For Rowlands, the intersubjective interaction is at the beginning of moral action, so Gallagher's emphasis on the second-personal perspective seems complementary.

Finally, when Gallagher remarks about the "interaction itself…contribut[ing] something that is not reducible to the actions of the individuals involved,"[5] the beginnings of what I have called the adaptivity criterion appear. Now, to be clear, Gallagher is not talking only about moral interactions, so the account of adaptivity I have discussed is more specific. But again, Gallagher's account seems consistent with the view I am defending.

But what of Gallagher's full empathy – the kind of empathy that depends upon both narrative and social competencies? It seems implausible to suggest that non-verbal NHAs who also lack linguistic abilities could have narrative competency. But perhaps there is conceptual space between Gallagher's direct perception capacity and his full empathy. What I mean to suggest here is that we might be able to pry apart narrative competency from at least some kinds of social competencies. If some NHAs demonstrate what I have called (moral) adaptivity that suggests the presence of shared social and cultural practices, then it seem such NHAs could fail to recognize themselves as narrative

[4] There may be an objection looming. One might worry that the behavior of an NHA does not prove that the NHA *feels* concern. And this is likely true. After all, granted that we lack access to the phenomenological experiences of NHAs, we cannot prove that they feel concern *per se*. However, we make these kinds of attributions – the attributions of feelings in the absence of linguistic articulation – consistently in the human case and with respect to pain or pleasure in the NHA case. I think then that making an attribution of concern-feeling to an NHA is not problematic, so long as (1) we acknowledge that we are only tracking an affect experienced by the NHA that may be different than our own experience of concern, and (2) we note that additional evidence for the claim that something like concern is experienced by NHAs is really an empirical question that I cannot resolve here.

[5] Gallagher (2012), 184.

beings while still demonstrating an awareness of and responsiveness to the intentions and circumstances of others in their moral community. Think, for example, of the complex grieving and burial rituals shared by African elephants, detailed in Douglas-Hamilton *et al.* (2006).

At this point, it is worth saying again that my goal here is to provide an account of empathy on offer in contemporary philosophy that is consistent with Rowlands's view. In no way is this chapter an exhaustive literature review of the varieties of theory of mind accounts of empathy that exist. This goal – to isolate one theory of mind account of empathy that is compatible with Rowlands's view – is part of the larger project of building upon Rowlands's account via the incorporation of empathy as a plausible, real-world equivalent for the conceptual moral module. This larger project is of interest for a variety of reasons, but perhaps most importantly, it provides us with a framework to employ as we take up the final challenge of Rowlands's book: going out into the world to evaluate empirical findings on NHAs to see if they in fact are mere moral subjects in the way he has described.

Revisiting Rowlands

To make more clear the *role* that I am suggesting empathy plays though, let me pause to revisit Rowlands's claims with respect to NHAs as moral subjects. Remember that he takes it that Myshkin stands as proxy for NHAs. All of the descriptions of Myshkin he provides are meant to serve the dialectical purpose of demonstrating how NHAs can be described (accurately) as acting for moral reasons. He suggests explicitly, that when NHAs do act for moral reasons, this is because of a kind of "fundamentally emotional" moral sensitivity.[6] When he provides specifics as to what emotions can be moral emotions (i.e., emotions that "possess identifiable moral content"), he lists the following as examples: "sympathy and compassion, kindness, tolerance, and patience, and also their negative counterparts such as anger, indignation, malice, and spite." He goes on to include also "a sense of what is fair and what is not."[7] All such emotions are joined together under his label of "concern."[8]

At this point, I want to make it clear that I am not quibbling with Rowlands about word choice, with his term "concern" and my term "empathy" capturing the same phenomena. For Rowlands, when an NHA is a moral subject, that NHA has an admittedly "psychologically unrealistic"[9] and intentionally ambiguous[10] moral module that reliably produces a normatively-assessable *and* fundamentally emotional sensitivity to the morally salient features of situations.[11] When he uses the term "concern," he is referring to a group of emotions (i.e., the one's that he has called "morally laden" emotions) that are produced by such a process. In short, concern is then the *output* of a mechanism, what he calls a "moral module." My suggestion is that empathy is the moral

[6] Rowlands (2012), 213.

[7] Ibid., 32.

[8] This move – that is, grouping all such emotions under the broader heading of "concern" – appears repeatedly throughout the book, but it is first made on p. 8.

[9] Rowlands (2012), 149.

[10] I mean that it is ambiguous in two respects. First, Rowlands uses the term to refer to "*whatever* mechanism plays the role of linking perceptions of situations with appropriate emotional responses" [emphasis mine]. And secondly, he does not wish to link his concept of the *moral module* to "any specific theories in cognitive science" [see p. 146, with respect to both quotes].

[11] Rowlands (2012), 230-231.

module, the means by which a mere moral subject's perceptions are linked to the morally salient features of the world.

This linking is experienced by the subject as a particular kind of sensitivity – one that arouses certain affective states that are strong enough to motivate the subject to act. When the subject in question both has these motivations reliably (i.e., when her affective states correspond to the morally salient features of the world with frequency) and is a member of a moral community (i.e., she is the kind of being who can be responsive to moral correction from her own interspecies or intraspecies community that shares moral practices), we observers can say of the mere moral subject that she acted for moral reasons. Empathy, in the way I am using the term, is then the actual mechanism that first makes possible moral sensitivity – and then makes ultimately possible displays of concern and more complex social practices that respond to the morally salient features of one's situation. To capture this range of roles that empathy plays in NHAs being mere moral subjects (when they are), I have employed Gallagher's language of direct perception and social competency.

A case for the moral subjecthood of rats

The members of a variety of species meet the sufficient conditions for being moral subjects, even if my adaptivity criterion is added to Rowlands's other three conditions. And since we now have an account of empathy to employ in defending the claim I have just made, I want to apply all of this to a test species, if you will. I have intentionally chosen a species that one might not expect, though there is similar evidence in many other species.[12]

[12] See Sanjida O'Connell (1995) regarding chimpanzees, Frans de Waal (2008) regarding cetaceans, Lucy Bates *et al.* (2008) regarding elephants – to give just a few examples.

Ben-Ami Bartal *et al*. (2011) conducted an experiment regarding empathy in rats. Empathy is therein defined as a type of pro-social behavior, referring to "actions that are intended to benefit another."[13] According to their definition, empathy requires "other-oriented emotional response elicited by and congruent with the perceived welfare of an individual in distress."[14] Empathy is distinct from contagion, according to these researchers, as the former demands of the agent that her "own distress is down-regulated, thus allowing empathically driven pro-social behavior."[15]

Ben-Ami Bartal *et al*. conceive of a simple experiment in which rats are presented with trapped conspecifics. Ultimately, the researchers conclude that "rats behave pro-socially when they perceive a conspecific experiencing psychological restraint stress, acting to end that distress through deliberate action."[16] They add: "Moreover, this behavior occurred in the absence of training or social reward, and even when in competition with highly palatable food."[17] Though they consider alternative explanations, the researchers argue that "the most parsimonious interpretation of the observed helping behavior is that rats free their cagemate in order to end distress, either their own or that of the trapped rat…This emotional motivation, arguably the rodent

[13] Inbal Ben-Ami Bartal, Jean Decety, and Peggy Mason, "Empathy and Pro-Social Behavior in Rats," *Science* 334, no. 6061 (Dec. 2011): 1427.

[14] Ibid.

[15] Ibid.

[16] Ibid., 1430.

[17] Ibid.

homolog of empathy, appears to drive the pro-social behavior observed in the present study."[18]

The definition of "empathy" employed by the researchers also implies that the subject need not be acting selflessly (e.g., the free rats might have experienced unpleasant emotions when hearing the caged conspecifics vocalize). It is enough, according to Ben-Ami Bartal *et al*., that the emotional motivation (selfless or not) manifested in intentional efforts to relieve the stress of another.

Up to this point, I have only summarized the observations and conclusions of Ben-Ami Bartal *et al*. It is then important to spend some time discussing why this evidence supports the view I am defending herein. First, the capacity for direct perception (e.g., sensitivity to the cagemate's vocalizations, wriggling) seems evident here. Such a case appears consistent with the direct perception model of empathy, wherein meaning is conveyed via the conspecific's vocalizations and bodily movements.

But, importantly, it is not just that the other's distress registers in the perception of the observer rat. Rather, in accordance with interaction theory's emphasis on enactive perception, the rat's perception of the distressed other is action-oriented in that it affords possibilities for action (e.g., freeing the caged rat) precisely because the actions of the caged rat are meaning-laden. If empathy is simply the passive perception of another's emotional states, then it does motivate one to act (e.g., in a pro-social manner). However, if the perception given by empathy is enactive (such that it affords action opportunities), then this can explain how empathy motivates one to act (or in Rowlands's language, this explains how empathy produces the kind of normative sensitivity that leads – at least sometimes – to moral action).

[18] Ibid.

And moreover, the rat's perception of the distressed other is embedded. On the interaction theory model, cognition is the product of dynamic interactions between subjects and their environments – following Varela, Thompson, & Rosch 1991; Thompson & Varela 2001; and later Gallagher (who I have already discussed at length). I have already suggested that an interactionist reading of this kind of data is preferable to the others I have discussed, as it does not require that the rat in this case have the cognitive capacities either for imagining herself in the position of the other or for (even minimal) mindreading.

But there is another reason for preferring the interaction theory model that seems most fitting in this empirical section. Thompson & Varela put the point nicely:

> Although it is often tacitly assumed that consciousness must 'supervene' entirely on internal neural states, it is far from clear how one is supposed to distinguish between 'internal' and 'external' states. Despite the philosophical fiction of a 'brain-in-a-vat', it is doubtful (even as a thought experiment) that one can 'peel away' the body and the environment as 'external' to the brain processes crucial for consciousness. The nervous system, the body and the environment are highly structured dynamical systems… [A] better conception of brain, body and environment would be as mutually embedded systems…[19]

There is then good reason, as a scientist and/or as a philosopher, to be skeptical of models that suggest a commitment to a distinction (between internal and external states) that seems no longer to cohere with the now well-documented empirical findings regarding the necessity of explicating phenomena via a dynamic systems approach (e.g., sensorimotor coupling). For this reason, IT's emphasis on embeddedness also makes it a preferable model for interpreting the behavior of rats I am discussing.

I am defending a reading of Ben-Ami Bartal *et al*. that is consistent with IT, but also accords with Rowlands's view. The fact that the rats in this experiment demonstrated

[19] Evan Thompson and Francisco Varela, "Radical embodiment: neural dynamics and consciousness," *TRENDS in Cognitive Sciences* 5, no. 10 (Oct. 2001): 422-423.

helping/freeing behavior consistently when presented with the relevant stimuli (e.g., the caged conspecific's wriggling or vocalizations) strengthens the interpretation I am advancing: that these rats seem to behave consistently in response to some morally salient features of their environments. The suffering of the caged conspecific leads to many of the free rats releasing their conspecifics. In so doing, the actions of (some of) the free rats track a moral proposition (e.g., "suffering is bad") – even if those rats cannot entertain or scrutinize that proposition. If we adopt Rowlands's arguments, we are then right in saying that the rats have acted morally – and that the world in which creatures behave in this way is a better world than the one in which they do not.

But note that this first experiment is limited in focus, and in so being, it does not provide a clear case for thinking that rats meet my added adaptivity criterion – that they are responsive to moral teaching and correction (or feedback more generally) from a moral community. For that, we must turn to another study.

In Ben-Ami Bartal *et al*. (2014), researchers performed a number of experiments. The first of these is now familiar; the researchers replicated earlier experiments showing that free rats would lease caged conspecifics in most cases. But what is new here is that the experiments were broadened to include caged rats who were not of the same strain as their free conspecifics. What researchers observed was that free rats were more likely to help strangers from a different strain if the free rats had been fostered by members of that strain than the free rats were to help conspecifics from their own strain.[20]

[20] Inbal Ben-Ami Bartal, David A. Rodgers, Maria Sol Bernandez Sarria, Jean Decety, and Peggy Mason, "Pro-social behavior in rats is modulated by social experience," *eLife* (2014): 2.

According to the researchers, "…social interaction with another rat, such as that occurring while rats live together, is critical to shaping pro-social motivation."[21] Citing Dungatkin (2002) as more evidence to this point, Ben-Ami Bartal *et al*. add: "…relying on social experience rather than genetic similarity for guiding pro-social behavior has an added value in that it allows animals to flexibly adapt to different circumstances."[22] And according to the researchers, this adaptivity suggests the presence of empathy in rats[23] and demonstrates the ways in which rats form socially defined groups, or "groups based on social experience" (such that "genetic similarity does not influence pro-social motivation").[24]

What I want to suggest, via this appeal to Ben-Ami Bartal *et al*. (2014), is that rats have the kind of adaptivity that I have argued should be added to Rowlands's sufficient conditions. That is, they learn over time to respond to the suffering of conspecifics differently than instinct or genetics might suggest they would (even in the absence of human training). In this case, rather than genetic similarities determining their reactions, or humans training the free rats to open doors, the relationships had by the free rats with members of another strain impacted dramatically which caged conspecifics they freed.

But even more to this point about moral adaptivity in rats, Rutte and Taborsky (2007) performed another set of experiments involving rats. Demonstrating what the researchers call "generalized reciprocity," rats were more likely to help conspecifics if

[21] Ibid., 4.

[22] Ibid., 7.

[23] Ibid., 9.

[24] Ibid., 10.

they themselves had been recipients of help in the past.[25] In this set of experiments, if the conspecific of the rat in question had pulled the lever to give her food, then she was more likely to do so for others conspecifics. Bekoff and Pierce provide this gloss of the findings: “[The rats] generously help an unknown rat obtain food if they themselves have benefited from the kindness of a stranger.”[26]

In light of these considerations about both empathy, moral adaptivity, and the way that particular interactions inform and shape behavior, it seems we can defend the claim that rats are better understood as moral subjects who can sometimes act for moral reasons than as moral patients who cannot engage with their communities in this way.

[25] Claudia Rutte and Michael Taborsky, “Generalized Reciprocity in Rats,” *PLoS BIOL* 5, no. 7 (2007): 1422.

[26] Marc Bekoff and Jessica Pierce, *Wild Justice: The Moral Lives of Animals* (Chicago: University of Chicago Press, 2009), 21.

6

Some Implications Considered

Up until this point, my emphasis has been on NHAs – and more particularly, on why we ought to think of them as moral subjects, building from Rowlands's account. But in this chapter, it might be of interest if we switch gears a bit and consider additional implications of adopting Rowlands's three-part schema (i.e., between patients, mere moral subjects, and moral subjects who act agentially).

I have already addressed some implications about non-NHA cases. I have argued that robots like Ollie might be best understood as falling outside of any of these categories, but that it remains possible that more complex systems like Samantha could in fact be moral subjects. My primary reason for differentiating between these two cases had to do with Samantha's ability to break the rules, as opposed to Ollie's inability to do so. In moral cases, that difference was highlighted in what I called the adaptivity criterion.

Moving then to NHAs, the account proposed by Rowlands and developed herein has been that at least some NHAs are mere moral subjects. More particularly, in my view, NHAs who meet Rowlands's conditions (i.e., those who possess some mechanism that makes possible reliable normative sensitivity) and who also evidence moral adaptivity (i.e., responsiveness to feedback received from one's moral community) are mere moral subjects. I have also suggested that the capacity for empathy is a testable, real-world equivalent of Rowlands's moral module mechanism, and that some NHAs (e.g., rats) possess this capacity.

What might be of interest now is to consider what all of this might mean for some of us human animals. After all, since Rowlands's view allows for agency to be a decision to decision, moment to moment descriptor (rather than a static moral status), there may be some questions about where various humans fit into the picture that is being advanced herein. Though there are certainly many more cases to be considered, I want to briefly examine what we might say of three in particular, assuming this view of subjecthood and agency is adopted.

The sociopath

The first of these is the sociopath. I first became intrigued with the idea of bringing the sociopath into this conversation as a means of giving a *prima facie* case for a link between empathy and moral status. Consider our ordinary intuitions about sociopaths. In popular literature, film, and conversation, the sociopath is often referred to as a human who lacks empathy[1], sometimes even to the degree of suggesting that the lack of this singular capacity is sufficient to argue that the sociopath is somehow "subhuman"[2] or profoundly other.

And at first, my thought was to argue that sociopaths are in fact not agents. This would of course entail that they are not responsible for their actions in the way that most adult humans are; that is, they are not morally responsible in virtue of the fact that they lack empathy. When we punish sociopaths for their actions, I thought, it is for prudential reasons. We are trying to protect our communities, even if we believe that sociopaths are the kinds of beings who could not have chosen differently in virtue of their lacking this crucial feature of human psychology. The sociopath, in lacking this capacity for empathy,

[1] Even the DSM-IV-TR emphasizes the lack of empathy in the clinical description of anti-social personality disorder [see p. 703].

[2] In an episode of *The Sopranos*, a character – in fact, a psychiatrist named Richard LaPenna – refers to Tony Soprano as a "sociopath," as "untreatable," and as "scum" in an initial conversation with LaPenna's ex-wife. In a later conversation in the same episode, LaPenna additionally refers to Soprano as "subhuman" [see Season 1, Episode 8].

might fairly be characterized as amoral – that is, as profoundly insensitive to the morally salient features of the world (e.g., the suffering of others). And note that even if this is right, there is no reason to think we should not understand sociopaths as moral patients (i.e., as objects of our concern who should not be treated in certain ways).

And to this view, note that it is not with reference to other deficits (i.e., besides the deficit for empathy) that we most often characterize the sociopath. When we consider popular depictions of sociopaths, it is not the lack of cognitive abilities that sets the sociopath apart. In fact, some sociopaths (e.g., Hannibal Lecter) are described as being exceptionally intelligent.

Nor is it the case that we talk about sociopaths as if they lack all capacity for affect. Sociopaths seem capable of presenting genuine excitement, anger, and sadness. Consider the depictions of the Joker's thrill at a plan unfolding as he intended or Norman Bates's anger in response to harms suffered by his mother. What sets the sociopath apart from the average adult human is that the sociopath does not experience the appropriate affect in morally significant situations (or, at least, that he/she does not experience appropriate affect *reliably*[3]).

To be sure, we occasionally employ other characterizations for sociopaths (e.g., narcissism). But it seems we could easily explicate a person's narcissistic tendencies as following from her lack of empathy. In fact, this is exactly the narrative given in the popular book *The Sociopath Next Door* (2005). Psychologist Martha Stout describes

[3] I add this qualification because one might think that Norman Bates is morally right to feel anger in response to his mother being attacked. But we would still not say of him that he is reliably right in this way.

sociopaths as “hav[ing] no trace of empathy,”[4] such that the sociopath can be said to have “no conscience.”[5] And to be clear, when she uses the terms empathy or conscience, she means “a sense of obligation based in an emotional attachment to another creature…”[6] This lack of emotional motivation to act in other-regarding ways (or empathy, more simply) is identified by her as the central feature of the sociopath, with all other behavioral outputs following from this deficit. Again, the capacity for empathy moves front and center. Over and over, whatever that *thing* is that makes sociopaths different from the bulk of us is called “empathy.”

My initial thoughts on the sociopath centered on this idea that empathy’s very presence (or lack thereof) seems to carry a tremendous amount of weight with respect to moral status – at least if we are inclined to think that the sociopath is not a moral agent (or subject more broadly). One reason to think this has been hinted at above – that is, maybe the sociopath has no *reliable* normative sensitivity, since she has no capacity for empathy.[7] But wait… it seems many sociopaths do have reliable normative sensitivity!

I now have in mind a kind of sociopath, but not the kind of sociopath to whom I previously referred – that is, not the easily identified, violent, hyper-intelligent sociopaths who make the news and fill television scripts. Instead, I am thinking of the peaceful sociopath, the kind who goes unnoticed.[8] These sociopaths are generally law-abiding.

[4] Martha Stout, *The Sociopath Next Door* (New York: Broadway Books, 2005), 7.

[5] Ibid., 25.

[6] Ibid.

[7] Recall the example of Norman Bates. While he may occasionally get it right, most of the time his normative sensitivity fails to produce in him the appropriate affect.

[8] According to Stout’s data, one in every twenty-five people is a sociopath, and most of these individuals are neither violent nor criminal [see p. 9].

Despite the deficit of empathy, these sociopaths live relatively normal lives. So in Rowlands's schema, since these sociopaths have a variety of interpersonal relationships and are generally law-abiding, it seems they evidence a normative sensitivity that reliably tracks the morally salient features of situations. After all, in an important way, many of their actions track moral features of the world (e.g., not stealing or murdering).

So what is it that they lack? Is the deficit correctly characterized as a deficit of empathy? I think there are two responses one could take, again assuming the view that has been defended herein. First, one might suggest that the answer is found in the belief that sociopaths are untreatable – or to use my language, that they are not adaptive in response to moral correction. That is, they cannot be inducted into moral practices via adaptivity because they do not care about others around them. If they are not by disposition already sensitive to some moral feature of the world, then they cannot be made to be so. I think this is a critical element of the way we conceptualize sociopathy and may speak to the plausibility of thinking that adaptivity (in addition to sensitivity) is critical for moral status. There has to be something in the subject that makes her able to act caringly with respect to the interests of others in a way that might not come naturally to her.

But there may be objectors to this presentation of sociopathy, though it is of course not intended to be comprehensive. One might highlight the role that accountability plays in moral agency *and* find it puzzling yet defensible that sociopaths are often punished for their crimes. Or there might be some who think this characterization of sociopaths as lacking adaptivity is unfair or incomplete. As I mentioned above, there is a second reply available here. It might be the case that sociopaths are correctly identified as agents – but not moral agents (i.e., moral subjects who can act agentially in virtue of their metacognitive abilities). There are, after all, others ways of being an agent. Perhaps the sociopath is not capable of acting for moral reasons, but she is capable of acting for reasons (e.g., merely prudential ones). And in virtue of her cognitive abilities (or something like that), perhaps we can assign legal responsibility to her *qua* being a legal agent, even if she is not a moral agent.[9]

Sociopathy is of course one of the hardest cases to consider here. Trying to figure out where to locate the sociopath, especially the more garden variety sociopath who is generally law-abiding, is tremendously difficult and requires a more detailed analysis than I can provide, granted the primary purpose of this project. And I should note that there is a third response available here, but it requires abandoning Rowlands's claim that moral subjects and agents are different in degree (but not in kind). One might think that sociopaths are in fact moral agents, but that there are perhaps two routes to agency – with one being more affective and the other more cognitive. But I will not elaborate on that idea here, as the purpose of this chapter is to offer some thoughts on how one might respond to certain human cases if one is also committed to Rowlands's framework.

The Kantian agent

The second case I want to consider here is a familiar account of moral action that would suggest that the dispassionate but cognitively sophisticated creature, who is motivated by her rational capacities alone to act for moral reasons, is actually the ideal

[9] Summaries of David Shoemaker's contribution to Thomas Schramme's *Being Amoral: Psychopathy and Moral Incapacity* (2014) suggest that Shoemaker makes an argument along these lines. In short, he argues that sociopath's have legal or criminal responsibility, even though they may not have moral responsibility.

moral agent. The account to which I refer is of course that of Immanuel Kant. There is a nice summary of Kant's views on moral motivation and emotion from Justin Oakley:

> A central claim of Kant's ethics is that only acts which are done from duty have moral worth. In the course of arguing for this claim Kant explicitly asserts that, apart from one important exception, motivation by emotion cannot be morally good. The emotion which Kant excludes in making this assertion is respect for the moral law, since in his view this is the only emotion which involves a recognition of the determination of the will by the moral law.[10]

In short then, the Kantian agent is motivated to act solely from duty, and duties follow from the employment of reason.

The supremacy of reason over emotion, with respect to moral action, centers on the former's relative reliability, according to Kant. Oakley summarizes: "Kant's main argument against emotions as moral motives in the *Foundations* [referring to Kant's *Groundwork of the Metaphysics of Morals*] is that emotions cannot be moral motives because they are unreliable and capricious, whereas duty is dependable and stable, and is always available to us as a motive to action."[11] It would of course follow that Rowlands's mere moral subject is in great trouble.

Rowlands has already given an argument as to why we ought to think the Kantian account is ultimately indefensible. That conclusion is the one he pursues by introducing the reflection condition and then undermining its assumed link between moral scrutiny and control. But we have another reason available in my view for thinking that the ideal Kantian agent is missing part of the picture.

[10] Justin Oakley, "A Critique of Kantian Arguments against Emotions as Moral Motives," in *History of Philosophy Quarterly* 7, no. 4 (Oct. 1990): 441.

[11] Ibid., 443.

To get at what I have in mind, let me begin by bringing in a criticism of Kant via Oakley. Oakley charges that the person who is motivated to act by moral emotions is in fact better able to fulfill Kant's duties than a purely rational person could. He explains, as follows:

> Now, an important reason why a sympathetic or compassionate person may be better able to carry out a duty of beneficence is that having sympathy or compassion for another is often necessary to gaining a proper understanding of what actually needs to be done in order to help her. But even apart from the superior insight into the situation which the sympathetic person seems likely to have, the help he provides to the recipient here may be more appropriate than that given by someone motivated by duty, because what the person in distress may need is action-from-sympathy, that is, action which is motivated by a feeling-with the other, rather than action done out of duty. Without such emotional motivation, one may fail to provide what the person in distress needs to alleviate her suffering, and so one may fail to carry out one's duty here. Thus, *there seems to be an inherent problem with motivation by duty which is unaccompanied by sympathy and compassion*, in the case of certain duties of beneficence; for, paradoxically, it is the very fact that one acts from duty uninformed by sympathy or compassion which entails that one fails to fulfill one's duty here [*italics added*].[12]

The argument then goes something like this: the Kantian agent cannot be the ideal agent, insofar as she does not rely on affect to guide her actions, because in relying only on reason, critically important information will be unavailable to her. That is, because she does not act from what Oakley calls "compassion" or "sympathy," she misses some of the morally salient features of her situation.

But even if one wants to maintain that the Kantian agent is still the ideal, in light of the fact that she will be sensitive to most of the morally salient features of a situation, this can be accommodated by Rowlands's view. My idea here is that a person cannot arrive at moral principles via reason alone – that is, without having first had affective states that motivated her to be concerned about others in the first place.

[12] Ibid., 454.

I am not the first to contend that empathy, understood as a mechanism that produces affective states (at least at first), has to precede the universalizing of a (rational) judgment. John Deigh makes a similar point. He uses the example of a woman who is motivated to care about the "freedom and well-being" of others. As he puts it, "…to arrive at this conclusion [i.e., that she ought to be concerned about the freedom and well-being of someone else] requires something besides applying the criterion of universalizability to her own judgment that others ought not to interfere with her freedom and well-being." He elaborates: "It requires instead empathy with the person whose freedom and well-being she judges that ought not to interfere."[13]

My hope here is that, in explicating Rowlands's moral module in terms of a capacity for empathy, a response like this is now available. Even if one is committed to the idea that an agent who acts from duty and reason is the ideal, we can still argue that empathy (and the affective states it makes possible) must have been present in the development of such an agent. It is just that this moral agent has employed her metacognitive abilities to ultimately suppress her affectivity.

The indoctrinated child

The last case I will consider in this chapter is that of the indoctrinated child. The relevance of this case may be the most difficult to see immediately. Albeit in another chapter, I have considered Ollie, a creature with no metacognition and no affect. In this chapter, I have considered the sociopath, who might be characterized as having limited or misguided affect but full metacognitive abilities. I have also considered the Kantian agent, who clearly has both affective and metacognitive abilities, even if the latter is employed to minimize her experiences of the former. The indoctrinated child stands in

[13] John Deigh, "Empathy and Universalizability," *Ethics* 105, no. 4. (July 1995): 758.

for the combination of abilities I have not yet engaged: that is, as a human who has affective capacities but not metacognition. If you like, this is a human instance of mere moral subjecthood.

Imagine a child who is developing typically. All I mean to suggest here is that her affective and cognitive skills are not deficient for her age, but she has not yet developed very many metacognitive skills. Imagine further that she is being raised in an environment wherein her guardians teach her to act in ways that go against her empathic inclinations. So even though she cannot reflect on her actions, she acts in ways that communicate hate (or something of that nature) to other humans. She is reliably sensitive to some of the moral facts (e.g., that the child she is teasing is sentient). In fact, the success of her bullying in part depends on the other child's being sentient – even if the bully does not recognize that consciously.

What do we make of this case? If the indoctrinated child is young enough that she has not yet developed metacognitive abilities, I think we will not want to suggest that she is morally blameworthy. Nor do I think we would be right to say of her that she lacks empathy or adaptivity. After all, her behaviors suggest that she is tracking some of the morally salient features of a situation (e.g., that the boy she is teasing is upset by it), and she has responded to instruction from members of her moral community (e.g., her guardians). For these reasons then, we should conclude that the indoctrinated child is a mere moral subject (and not simply a patient).

If I were to stop here, then the indoctrinated child would just be one of many instances of humans who are mere moral subjects. But her introduction, and the reason I have described her as a bully, serves an additional purpose: making sense of the last of

Rowlands's characters, namely Bizarro Myshkin. Bizarro Myshkin "delights in the suffering of others and feels repulsion at their happiness."[14] Rowlands describes Bizarro Myshkin as possessing a reliable moral module and as "morally evil."[15] Yet, presumably, Bizarro Myshkin, like Myshkin, lacks metacognition. So when Rowlands uses the term "evil," he cannot mean that Bizarro Myshkin is morally blameworthy (since such a creature is not an agent in his account). Imagining the indoctrinated child, who bullies in light of what she has been taught, gives us a real world example of Bizarro Myshkin and a human instance of mere moral subjecthood.[16]

[14] Rowlands (2012), 231, footnote 13.

[15] Ibid.

[16] This case also raises a question about what form(s) moral induction might take. I am implicitly allowing for brainwashing to count, in additional to other ways of being inducted into a moral practice (e.g., behavioral training, rational persuasion).

7

Discussion, Conclusions, and Recommendations

Starting with an account on offer from Rowlands regarding how it is that NHAs can act for moral reasons, I sought to expand on his view. After detailing his argument in Ch. II, I reflected on a few open questions that followed from careful consideration of his work. In Ch. III, I responded to these and proposed that his account of the mere moral subject might benefit from the addition of another condition (i.e., the adaptivity criterion). I then went on to introduce the idea that a plausible and psychologically realistic "moral module" could be empathy. In Ch. IV, I explored the complexities of employing the term "empathy," with a focus on contemporary theory of mind debates about this capacity. After outlining some of the various ways in which empathy has been characterized, I argued in Ch. V that Gallagher's interaction theory account was most helpful for my purposes – that is, for accounting for the varieties of empathy that (at least some) NHAs possess. Finally, in Ch. VI, I considered what adopting Rowlands's account of moral status might mean for certain human cases.

To begin my concluding remarks, I would like to go back to where this dissertation started – with the title. The promise of that title was to offer an account in which empathy plays a critical role in determining the moral status of NHAs. To this end, I have argued that empathy is a real-world mechanism that can link perceptions to the morally salient facts of a situation. If empathy can do this, and we can demonstrate evidence of empathy in NHAs, then those NHAs meet Rowlands's conditions for moral subjecthood.

But I also argued that another condition needed to be added to Rowlands's list: what I called the adaptivity criterion. Since I have not made this explicit, I need to do so now. Adaptivity in NHAs (i.e., being responsive to feedback from one's moral community) – just like the normative sensitivity that preexists it – requires a capacity for empathy. In the absence of language, there must be a mechanism present that allows an NHA to be sensitive to feedback.

The interaction theory model of empathy I have discussed herein offers a plausible way of thinking about how this feedback might be transferred between beings who might lack certain cognitive faculties or any linguistic abilities. Recall the description of Gopnik & Meltzoff (1997) offered by Gallagher. The researchers refer to infant's ability to "tune" their behaviors to those of another person, suggesting, as Gallagher puts it, that "…it is the interaction itself that contributes something that is not reducible to the actions of the individuals involved."[1] I am suggesting that something analogous, a similar kind of attunement, occurs in NHAs who are inducted into moral practices. As the infant via her interactions with another attunes her behavior, so too does the rat via her interactions with the conspecific who provided her with food and the one now desirous of receiving food from her.

As I am deeply sympathetic to the idea that NHAs are moral subjects, developing Rowlands's account has been something of a labor of love. By this, I mean to indicate not only that I am genuinely invested in his project, but also that I think we have a need for an account that makes sense of the moral lives of NHAs. Trapped between the binary of moral patienthood and agency, we have too long been left unable to account not only for

[1] Gallagher (2012), 182.

the ample number of anecdotes available regarding the moral interactions of NHAs with conspecifics and with humans, but also the growing body of empirical literature on empathy, caregiving, pro-social behavior, and altruism in other species. In light of the import – perhaps even the necessity – of producing a philosophically rigorous account of moral status that can accommodate what we have long known and what are learning about NHAs, my appreciation for Rowlands's recent book can hardly be overstated.

In light of this appreciation, I have attempted to expand upon Rowlands's account of the moral subject and to offer plausible ways of responding to questions or worries that some might have. But I would be remiss, perhaps, if I did not mention one worry that I have not been able to resolve for myself as of yet. My concern is that this model of mere moral subjecthood and its relationship to agency implies a kind of moral hierarchy, wherein humans alone can land at the far end of the spectrum pictured in Figure 1.

To be fair, I am not sure that this is a worry for Rowlands or for most philosophers. Perhaps the idea that humans are uniquely capable of detecting most of the morally salient facts in the world accords with what one believes to be true about our capacities (as opposed to member of other species). But I am in general skeptical about these kinds of depictions – those where the human way of doing some thing is depicted as the golden standard. But more to the point with respect to this particular project, I wonder if NHAs are actually better in many cases at being sensitive to the moral facts of a situation than we humans are. It seems to me that often our cognitive capacities and abilities to (over)scrutinize get in the way of our moral sensitivity, undermining at least in some of us a more immediate pull to care for the other.

This remaining worry relates to an area of future inquiry that I believe might be productive. Care ethicists echo some of the worries I have about thinking of morality in overly cognitive ways. Perhaps an engagement of this account with care ethics could produce a picture of moral subjecthood that could do even more justice to the moral lives of NHAs.

Even if the view I have advanced is considered, there is much empirical research to be weighed and conducted. I have only given a few cases for consideration here, but I think we could find evidence of empathy and moral adaptivity in many more species.

For these reasons, my interest in and work on this topic is hardly complete. Moreover, I hope that as more and more philosophers are introduced to Rowlands's recent work, they will come to share my interest in building from his framework. As philosophers, I believe that we have an obligation to make sense of the truths we come to see and to offer theories that reflect them. As ethicists, I believe we have an obligation to offer a voice for those who do not have one – whether that voicelessness be the product of institutional injustice, disability, or species membership. This dissertation is of course an ultimately small effort towards those goals, but I hope that it will be received as an effort towards them nonetheless.

References

Ben-Ami Bartal, Inbal, Jean Decety, and Peggy Mason. "Empathy and Pro-Social Behavior in Rats." *Science* 334, no. 6061 (Dec. 2011): 1427-1430.

Ben-Ami Bartal, Inbal, David A. Rodgers, Maria Sol Bernandez Sarria, Jean Decety, and Peggy Mason. "Pro-social behavior in rats is modulated by social experience." *eLife* (2014): 1-16.

Bates, Lucy A., Phyllis C. Lee, Norah Njiraini, Joyce H. Poole, Katito Sayialel, Soila Sayialel, Cynthia J. Moss, and Richard W. Byrne. "Do Elephants Show Empathy?" *Journal of Consciousness Studies* 15, no. 10-11 (2008): 204-225.

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

Clark, Stephen. "Good Dogs and Other Animals." In *In Defense of Animals*, edited by Peter Singer, 41-51. New York: Basil Blackwell, 1985.

Darwall, Stephen. *The Second-Person Standpoint*. Cambridge, MA: Harvard University Press, 2006.

De Waal, Frans. "Putting the Altruism Back into Altruism: The Evolution of Empathy." *Annual Review of Psychology* 59 (2008): 279-300.

Debes, Remy. "From Einfühlung to Empathy." Invited chapter to *Sympathy*, part of the *Oxford Philosophical Concepts* series. Edited by Eric Schliesser (forthcoming).

Deigh, John "Empathy and Universalizability.*"* *Ethics* 105, no. 4. (July 1995): 743-763.

Descartes, René. *Selected Philosophical Writings*, 1988 reprint, translated by John Cottingham, Robert Stoothoff, and Dugald Murdoch. Cambridge: Cambridge University Press, 1999.

Diagnostic and Statistical Manual of Mental Disorders, fourth edition: DSM-IV-TR. Washington, DC: American Psychiatric Association, 2000.

Dillard-Wright, David. "Sympathy and the non-Human: Max Scheler's Phenomenology of Interrelation." *Indo-Pacific Journal of Phenomenology* 7, no. 2 (Sept. 2007): 1-9.

Gallagher, Shaun. "Neurons, neonates and narrative: From empathic resonance to empathic understanding." In *Moving Ourselves, Moving Others, Motion and Emotion in Intersubjectivity, Consciousness and Language*, edited by A. Foolen, U.M. Lüdtke, T.P. Racine, and J. Zlatev, 167-196. Amsterdam: John Benjamins Publishing Company, 2012.

Gruen, Lori. *Ethics and Animals: An Introduction.* Cambridge: Cambridge University Press, 2011.

Haraway, Donna. *The Companion Species Manifesto: Dogs, People, and Significant Otherness.* Chicago: Prickly Paradigm Press, 2003.

Hume, David. *A Treatise of Human Nature: The Clarendon Edition of the Works of David Hume.* Edited by D.F. Norton and M.J. Norton. New York: Oxford University Press, 2007.

Jahoda, Gustav. "Theodor Lipps and the Shift from Sympathy to Empathy." *Journal of the History of the Behavioral Sciences* 41, no. 2 (2005): 151-163.

Nichols, Shaun. "Mindreading and the Cognitive Architecture Underlying Altruistic Motivation." *Mind & Language* 16, no. 4 (Sept. 2001): 425-455.

Oakley, Justin. "A Critique of Kantian Arguments against Emotions as Moral Motives." *History of Philosophy Quarterly* 7, no. 4 (Oct. 1990): 441-459.

O'Connell, Sanjida. "Empathy in Chimpanzees: Evidence for Theory of Mind?" *Primates* 36, no. 3 (July 1995): 397-410.

Regan, Tom. *The Case for Animal Rights*, 1983 reprint. Berkeley: University of California Press, 2004.

Rowlands, Mark. *Can Animals be Moral?* New York: Oxford University Press, 2012.

Rutte, Claudia and Michael Taborsky. "Generalized Reciprocity in Rats." *PLoS BIOL*, 5, no. 7 (2007), 1421-1425.

Scheler, Max. *The Nature of Sympathy*, 2008 reprint. New Brunswick: Transaction Publishers, 2011.

Shoemaker, David. "Psychopathy, Responsibility, and the Moral/Conventional Distinction." In *Being Amoral: Psychopathy and Moral Incapacity*, edited by Thomas Schramme, 247-274. Cambridge, MA: MIT Press, 2014.

Smith, Adam. *The Theory of Moral Sentiments*. Edited by D. Stewart. London: Henry G. Bohn, 1853.

Stein, Edith. *On the Problem of Empathy*, third edition. Translated by Waltraut Stein. Washington D.C.: ICS Publications, 1989.

Stout, Martha. *The Sociopath Next Door.* New York: Broadway Books, 2005.

Stueber, Karsten. *Rediscovering Empathy: Agency, folk psychology, and the human sciences.* Cambridge, MA: MIT Press, 2006.

Thompson, Evan and Francisco Varela. "Radical embodiment: neural dynamics and consciousness." *TRENDS in Cognitive Sciences* 5, no. 10 (Oct. 2001): 418-425.

Titchener, Edward B. *Lectures on the Experimental Psychology of the Thought-Processes.* New York: The Macmillan Company, 1909.

Wittgenstein, Ludwig. *Philosophical Investigations*, fourth edition. West Sussex: Wiley-Blackwell, 2009.

Vischer, Robert. *On the Optical Sense of Form: A Contribution to Aesthetics.* In *Empathy, Form, and Space: Problems in German Aesthetics*, edited by Harry Francis Mallgrave and Eleftherios Ikonomou, 89-123. Santa Monica, CA: Getty Center, 1994.

Zahavi, Dan. "Max Scheler." In *History of Continental Philosophy III*, edited by A. Schrift, 171-186. Edinburgh: Acumen Press, 2010.