An Evolutionary Vindication of Morality

How to do evolutionary ethics without debunking moral normativity

Reber Roald Iversen

Thesis presented for the degree of

MASTER OF PHILOSOPHY

Supervised by Professor Bjørn Torgrim Ramberg

Department of Philosophy, Classics, History of Art and Ideas

UNIVERSITETET I OSLO

June 2019

Acknowledgements:

Firstly, thanks to my supervisor Bjørn Torgrim Ramberg for excellent philosophical guidance; from my initial thesis proposal about some methodological topic in political philosophy, to arriving at the core issue I was interested in. Secondly, I would like to show my appreciation for fellow philosophy student and friend Paal Fredrik Kvarberg. Not only for reading my drafts, but also for our inspirational discussions throughout the years. Finally, my heart goes to Marianne for being incredibly helpful and understanding during the final writing process.

Abstract:

The thesis has three main tasks. First, establish an evolutionary basis for how morality may have evolved. Second, refute that this basis ultimately ends up debunking morality. Third, develop a strategy for showing how evolution might be used to vindicate our moral practices. The first chapter seeks to establish why natural selection may have opted for prosocial traits in the course of biological evolution and why these traits – thought important for the evolutionary origins of morality – comes short of bona fide morality. Chapter two starts of arguing that morality certainly is adaptive but then argues at length against moral nativism – the thesis that the evolution of morality should be explained by reference to a genotype causing the moral trait to emerge. I end the chapter with a discussion of how cultural evolution might have shaped our social cognition. In the third and final chapter, I present an alternative genealogy of morality based on cultural evolution. I use this approach to show how a vindication of morality is available by considering the inter-subjective nature of cultural information. I end by outlining the sense in which moral normativity in terms of practical reasons are located independently of the psychology of the individual, but dependent on human psychology in general.

Contents

Introduction: The evolutionary debunking of morality

The idea that evolutionary theory can undermine the justification some of our beliefs goes back to Darwin himself.[1] Darwin was a devout Christian and he worried for what his theory of evolution might mean for his own religious beliefs. He worried that his beliefs about evolution *debunked* his beliefs about God. For if the human mind is a product of evolution by natural selection like that of other animals, then our beliefs in God could, in principle, be explained by their usefulness for human reproductive fitness, and not by their truth. Some of Darwin's contemporaries worried that his theory inevitably undermined morality to the extent of dragging normative ethics into a slump. In *The Descent of Man*, Darwin did develop a detailed account of how a human "moral sense" might have evolved and what kind of natural material morality consisted of. But if morality has *evolved* by something like the principles of natural selection and not by rational reflection and moral progression, "*the consequences would be disastrous indeed! We should be logically compelled to acquiesce in the vociferations of [those] who would banish altogether the senseless words 'duty' and 'merit'*" (Mivart 1871/2008, p. 204).[2] Although he worried for religion, Darwin remained unmoved by the alarmist warnings about the status of morality. But ever since the rise of Darwinism some 150 years ago, the thought that evolutionary theory might undermine morality by debunking our moral beliefs has never vanished. In fact, the last decades have seen a range of influential publications sustaining the evolutionary debunking argument.

In which sense does evolution allegedly debunk morality? Different evolutionary debunking theorists have slightly different answers to this question depending on the target of their argument, but there is a common denominator to all of them. All seek to undermine morality in some sense by alleging facts about its origin and all adopt the framework of Darwinian evolution to explain that origin. The main idea is that if our "evaluative attitudes" such as morality is "heavily saturated by evolutionary forces" (Street 2006, p. 114), the contribution of evolution might explain why our evaluative attitudes such as morality are the way they are. Furthermore, if it turns out that the evolutionary explanation actually contradicts some features

[1] Debunking arguments, however, have been around at least since fourth century BC (See Joyce 2016, ch. 7 for discussion)

[2] George Mivart is famous for initially being a stark defender of natural selection who later became one of its fiercest critics. He tried to reconcile Darwin's theory with the view of the Catholic Church but ended up being rejected by both.

that we initially attribute to morality, then the evolutionary explanation *debunks* those attributes. Therefore, it *explains away* what we take to be central and important to our understanding of morality, leaving no room for what we take to be the proper explanation of those attributes. But since we have different theories about what morality is and what its distinctive features is, we have different theories about which element of morality vulnerable to debunking. For example, it has been argued that the notion of a moral fact flies in the face of evolution (see Mackie 1977). Some debunking theorists claim that evolutionary theory debunks moral realism by attacking the notions of moral objectivity (see Ruse 2009 & Street 2006). Others claim that research on the evolution of certain common moral intuitions casts their practical relevance into doubt (see Singer 2005 and Greene 2008). Finally, there are people who argue that the practical authority of moral reasons, or moral normativity, in general is undermined by evolutionary theory (see Joyce 2006; 2016). Accordingly, with the help of debunking theorist Richard Joyce (2016, p. 143) we can distinguish different types of evolutionary debunking arguments:

(1) All our moral judgments are false.
(2) All our moral judgments are false in so far as they involve claims to objectivity.
(3) Some normative theories (Kantianism most prominently) based on certain types of moral reasoning lack justification.
(4) All our moral judgments lack relevant justification.
(5) All our moral judgments lack justification and permanently so.

We can see that the scope of the different arguments varies greatly with it their plausibility. For my purposes, only one of these will be relevant. (1) is a consequence of the full-blown error-theoretic view that evolution renders the notion of a moral fact a myth. (2) is the consequence of the evolutionary debunking of moral realism. (3) is the claim that certain normative theories based on certain intuitions draw their appeal from conformity with these intuitions, and therefore to undermine this type of intuitions is to undermine this type of theories. In this thesis, I will focus on (4). This type of evolutionary debunking argument is supposed to work as an objection to moral judgments, not necessarily that they are false, but that they lack relevant justification. But what is meant by relevant justification? The claim is that a necessary part of being a moral judgment is that they imply a special sort of normativity – moral normativity.[3] This is what makes them different from other types of judgments. In this case, the debunker

[3] What I mean by moral normativity will be developed throughout the thesis but will be explicitly discussed in sections 1.4 and 3.2-3.4.

claims that, behaviorally, the trait is manifest (i.e. we act as if morality normativity exists), but that this is due to evolutionary forces. If this is true, the debunker claims, then morality lacks relevant justification. For the evolutionary forces responsible for causing the trait to emerge isn't a *normative* process but rather *causal*. However, proponents of (4) usually formulate this objection as a challenge and not an a priori truth, because it depends on the details of the evolutionary account and the analysis of moral judgment they rely on. Therefore, their arguments don't go so far is to imply (5). Its more like moral agnosticism than moral atheism.

A prominent version of the evolutionary debunking argument (4) is defended by philosopher Richard Joyce. In his wonderful book *The Evolution of Morality* he develops the hypothesis that human beings are furnished with a "moral sense" bred by natural selection in the course of evolution. This moral sense is an *innate faculty that should be explained by reference to a genotype causing the trait to emerge*. It follows that manifest morality is to be explained by this evolved moral sense. Joyce thinks that, in the sense that the moral sense is real, morality does indeed exist. Morality is different from other types of normative thinking, and our thinking is influenced by morality. More to the point, we are moralizing creatures – we constantly walk around judging this to be wrong and that to be good, all from a moral perspective. Evidence from psychology and neuroscience confirms this. For example, human beings as young as three years old are able to distinguish moral rules from conventional rules, and moral transgressions from conventional transgressions (Smetana 1989; Turiel 1993). fMRI studies suggest that there is distinct activity in the brain's emotional centers connected to the moral realm as opposed to other domains. When subject was elicited to make moral judgments, distinct neural activity was observed (Greene and Haidt 2002). What are the features that distinguish morality? The most striking and unique feature of morality, Joyce observes, is that moral judgments seem to have a kind of inescapability. The reasons implied by such judgments seem to hold irrespective of individual's existing desires or motivations. However, such reasons seem to have a sort of practical authority as well; it must be possible to act on those reasons. Thus, when recognizing facts about moral reasons that should by itself produce motivation to act in accordance with it. But since Joyce holds that the emergence of the "trait" in question is caused by a genotype forged primarily by natural selection, we must ask whether natural selection would be likely to equip the moral sense with the ability to track moral facts about practical reasons. Joyce's answer is in the negative. He therefore provides a story about why we actually do morally judge, but that all those judgments are unjustified from the moral perspective. This is the crust of his evolutionary debunking argument.

But why assume that an innate moral sense wouldn't be equipped by natural selection to track moral facts? Our theory of the natural selection of our visual abilities, for example, must include the assumption that the capacity to see involved tracking facts about the environment. A very long time ago, *"…organisms that developed receptor cells housing photopigments were able to transduce the variable photon bombardments confronting them from the external environment. This ability provided those organisms with valuable information about the environment and conferred on them an adaptive advantage over their blind rivals"* (Clyne 2015, p. 231). However, an investigation into the evolutionary origins of morality leaves us with a story that doesn't need to invoke the notion of a fact- or truth-tracking ability to account for why the moral sense increased the fitness of the bearers of the genotype. The overwhelming consensus among evolutionary approaches to the origins of morality is that morality evolved as a reaction to ensure social cooperation. Cooperation seems to be one of human being's biggest success-stories, and we see that in other species such as the social insects, cooperation is extremely fitness increasing and very likely to be selected for by natural selection. If morality evolved for the sake of social cooperation, however, the moral sense doesn't have to be fact- or truth-tracking to be fitness advancing. In fact, Joyce thinks that the *appearance* of tracking some moral fact about practical reasons is better explained by what natural selection might have done to our motivational systems. Although beliefs aren't our focus here, we can apply this to moral beliefs to see more clearly the debunking effect this creates. If our capacity for moral judgment didn't evolve by natural selection to track moral facts, in so far as our moral beliefs are generated by our moral judgments, what are the consequences for the epistemological status of those beliefs? Of course, if this is true, then proper justification for our moral beliefs seem to be undermined.

Guy Kahane (2010, p. 106) has outlined a general form such an evolutionary debunking argument might take:

> *"Causal premise. S's belief that p is explained by X*
> *Epistemic premise. X is an off-track process*
> *Therefore*
> *S's belief that p is unjustified"*

An ‘off-track’ process is one that does not track truth or facts: it produces beliefs in a way that is not sensitive to whether those beliefs are true or represents the relevant facts. A version of this general argument can be found in many of the approaches of the evolutionary debunkers mentioned above including that of Joyce’s. While Joyce’s evolutionary debunking project on the whole does involve detailed discussions of the debunking of our moral beliefs, I will constrain my discussion to that of the moral judgment. Of course, the moral judgment is central to forming moral beliefs, but require additional argument and discussion that I won’t be able to provide here. Thus, my focus will be on the evolutionary vindication of our moral judgment. Of course, even though our moral judgment might be vindicated, there might be some intermediate process in between the judgment and the belief that may be susceptible to a debunking explanation, but that won’t concern me here. Joyce’s main argument for the debunking of our moral beliefs is that our capacity for moral judgment can’t be said to be tracking moral facts about practical reasons. Joyce thinks that our moral judgments imply a kind of normativity that we cannot account for in light of evolutionary theory. According to Joyce, we make a fundamental error when we make moral judgments, because in doing so we are presupposing a normativity that doesn’t exist. But that error has a use, therefore Joyce suggest, moral normativity may be an adaptive illusion. This is the claim I will concern myself with. An evolutionary vindication of morality must thus show how our moral judgment actually may imply such normativity in light of evolutionary theory.

There have been different approaches to vindicate morality in response to evolutionary debunking arguments. Prima facie, the most straightforward way is to flatly deny the epistemic premise in Kahane’s schema – that evolution by natural selection is not a necessarily an off-track process. This “philosophical” approach to the problem has been pursued by a number of moral realists (For examples of such responses, see FitzPatrick 2014; Shafer-Landau 2012). Typical for such responses is a firm distinction between the “descriptive side of ethics” and the “normative side of ethics”. They claim that even though you have some descriptive scientific explanation of the evolution of morality cast in causal terms, doesn’t have any serious consequences for the truth of normative propositions. This is because normative propositions can only be explained by a *reason-giving* explanation, which is distinct from a causal explanation (FitzPatrick 2014). Mixing the two, it is claimed, will inevitably lead to making the logical fallacy of inferring an “ought” from an “is”. Without going deeply into this debate, I am going to accept that the correct descriptive story about our moral capacities can have consequences for normative ethics. In this I am agreeing with a general theme the philosopher

Owen Flanagan (amongst others) has advocated for some time. Flanagan stress the way that paradigmatic normative theories, put forward by leading figures in the history of philosophy, have all provided a necessary descriptive counterpart theory about the relevant capacities needed to handle moral normativity. Examples include Kant's psychological faculty of pure practical reason, Aristotelian virtues as psychological dispositions, Humean moral sensibility and we can even include Peter Singers (2005) more "recent" discussion on intuitions. These assumptions have been – and continue to be – central to the substantive normative proposals favored by these theorists. They are also testable, and indeed many of them have been tested (see Flanagan 1991 for discussion and references). Followers of this naturalistic approach have called themselves *methodological naturalists*, and it is my interpretation that they have the common project of transferring the Quinean program of "epistemology naturalized" to the domain of ethics.

In this empirical spirit, my strategy for vindicating morality in light of evolutionary theory will be to attack the causal premise in the debunking argument, and since this premise precedes the epistemic premise, changing it will in turn have consequences for the epistemic premise. The causal premise in Joyce (and most debunkers) argument is the thesis of *moral nativism*. Joyce claims that the distinct *moral* contribution to moral judgments comes from 'within' and is best explained by reference to a genotype forged by biological natural selection. Therefore, the purportedly special type of normativity involved in moral judgments is *innately specified* information. Nativism about the moral judgment, therefore, is basically the first premise in Joyce's argument. Consequently, the question of whether we should think that the moral judgment involves some biological adaptation or not will take up a lot of space in this thesis. In fact, almost all of chapter two is devoted to this question. In short, my answer will be that although innate mechanisms are necessary for the moral judgment, they are not sufficient. However, by focusing on the causal premise, I am still pursuing an *evolutionary* vindication of morality. Thus, I must explain how that which is not sufficiently explained by nativism – that which isn't innately specified information – also can be explained in light of evolutionary theory. For my purposes, this will mean outlining how an evolutionary approach can vindicate moral normativity from a moral perspective.

By pursuing this strategy, I agree with Joyce that morality does appear to imply a special sort of normativity and that any vindication must account for it. Additionally, it also means that my approach to evolutionary vindication of morality differs from the attempts of some fellow

methodological naturalists. It seems as if methodological naturalists are prone to find *instrumental* justification for morality. Daniel Dennett "god-trick" hypothesis is symptomatic of this (see Dennett 1995). As I see it, this leaves morality unvindicated. According to this strategy, morality is justified based on it being practically useful. But as I see it, this would still allow for morality to be massively mistaken, and thus leave morality unvindicated in the proper sense. Consider the case of religion. Religion cannot be vindicated in the relevant sense by it being useful to religious people. Presumably, a true believer would not agree that her faith is properly justified on the basis of being means to some other more valuable end. She would probably find it quite patronizing and disrespectful to be accused of being religious for such reasons. The same story holds for morality – we must be able to justify our moral judgments from the moral perspective.

To sum up, then. On two large starting points, I agree with Joyce. First, we are both methodological naturalists holding that our ethics should be informed by our best relevant science. For our purposes, this means a Darwinian or evolutionary approach. Secondly, we both agree that there is a descriptive sense in which morality does exist. Moreover, this empirical sense of morality suggests a conceptual understanding of moral judgments. There seem to be a special sort of normativity connected with moral judgments that distinguish them from other types of judgments. Chapter one will be devoted to developing these two points. Therein I will develop how natural selection might favor cooperative traits, but I also try to highlight why those traits are not moral traits without yet concluding whether the moral trait also is due to natural selection. That question will, as I mentioned above, be posed in the second chapter, where I maintain that moral nativism doesn't hold up to scrutiny. I end the chapter by suggesting that another evolutionary process – cultural evolution more precisely – may be relevant for explaining some of our more 'sophisticated' cooperative traits. In the third and final chapter I present an alternative genealogy of morality as opposed to the nativist genealogy. Finally, this anti-nativist proposal allows for an explanation of moral normativity that successfully vindicates our moral judgment in the relevant sense.

Before starting I would like to point out that throughout the thesis I will make use of empirical research from fields such as biology, anthropology, primatology, psychology and neuroscience. My familiarity with of this literature is not due to any particular scholarly relationship with those disciplines. My knowledge is based on reading some of my favorite philosophers operating in the debate on the evolutionary origins of morality and their treatment of the

empirical research. Reading them over the years have led me to a large interdisciplinary literature which I have tried to utilize in my own 'autonomous' way. But this is no guarantee that my treatment is free from bias and it should be expected that my overview of the all the relevant research is at best only partial. Another important thing to note is that the philosophy books and articles I have read on the subject were largely published between 2005 and 2012, and therefore the research they use date further back, meaning that there is a possible time-lag that might affect my arguments. A lot has happened in the field recently and I haven't acquainted myself enough with the contemporary cutting-edge literature to be comfortable using them herein. Therefore, I ask the reader to bear this in mind when considering the strength of my arguments.

Chapter one: The difference between altruism and morality

1.1 The possibility of altruism

I would like to start off with the claim that one of the most important subject matters of morality is *interpersonal relations*. A large body of extensive cross-cultural research have consistently been identifying some common denominators in moral systems: i) disapproval of certain acts of harming others, ii) values and norms related to reciprocation and fairness, iii) demands on appropriate behavior in relation to one's social status, and iiii) regulations of bodily affairs such as sexuality (see Haidt & Joseph 2004 for discussion and references). If we remind ourselves that what seems self-regarding from a Western perspective isn't necessarily the case for the rest of the world (here I am thinking of the fourth category), all four categories could be said to involve interpersonal relations. In fact, some empirical evidence shows that a considerable amount of the moral domain is used to determine how humans may harm one another (See Nichols 2004, ch. 7). This should lead us to suppose that morality is largely in place to facilitate and nurture the social order. Indeed, a key theme in this thesis will be to show that morality functions to ensure social cooperation. However, one might think that this flies in the face of evolution by natural selection. Natural selection, as one popular conception goes, favors selfishness. In this section, I will argue that this is a misplaced conception. In recent decades, there has been great progress in showing how cooperative traits[4] may have evolved. In so far as these traits are crucial to how we became creatures who morally judges – perhaps we can even say that these traits are the earliest springs of morality – it's natural to begin my argument by showing how natural selection may have installed cooperative traits in human beings. Great progress has been made in recent decades in theorizing different mechanisms for the evolution of cooperative traits and special attention has been given to the case for altruism. In the first part of this chapter I will focus on the evolution cooperative traits with special emphasis on the paradigmatic case of altruism before turning to explain how these traits are parts of the 'building blocks of morality',[5] but not the real McCoy. The concept of altruism is used in many circles both conventional and scholarly making it hard to get clear on what is meant by the term.

[4] I will use the word *trait* to mean a *behavioral* characteristic or quality (a phenotypic trait).
[5] A usefully vague phrase owed to the renowned primatologist Frans de Waal that will also figure in later discussion.

Luckily, philosopher of biology and evolutionary ethicist Philip Kitcher (2011, p. 19) gives us some useful distinctions to get us started:

(1) *Biological altruism*: Often called 'evolutionary altruism' and means behaving in a way that advances another individual's reproductive fitness, at the expense of one's own reproductive fitness. Contrast: fitness advancing.
(2) *Behavioral altruism*: Behaving in a way that benefits another individual: Contrast: harmful behavior.
(3) *Psychological Altruism*: Acting with the intention of benefitting another individual, where this is motivated by a non-instrumental concern for his or her welfare. Contrast: selfishness.

Biological and behavioral altruism is only relevant for us here to the extent that they work as contrasting notions – the crucial concept is that of psychological altruism. Psychological altruism has nothing to do with the spread of genes, but everything to with the intention of the agent. A psychological altruist is motivated by deliberative considerations oriented toward the welfare of others. In this sense, an act is altruistic only if the welfare of another was the agent's ultimate reason for acting. Thus, that which determines if the action is of the relevant sort is the motivations on which the agent is acting. If the agent is helping another, but the helping behavior is ultimately motivated by a deliberative belief that the helping will benefit the agent herself in the long run, this behavior is properly understood as selfish. Acting on a deliberative selfish motivating reason is a chronic feature of human beings in all societies and most certainly all societies that ever where. However, even for its precariousness, psychological altruism is a persistent achievement on (at least) an equal scale. I am confident to make the claim that the fact that people are often motivated to act on considerations for another's welfare is a commonsense observation, just as selfish actions are. That human beings have strong proclivities to be selfish is 100 % compatible with the possibility of psychological altruism. What I need for the argument of the thesis is only that psychological altruism is a real possibility. On this view there is room for selfishness. However, on the other extreme, there are those that believe that psychological altruism doesn't exist at all – in fact, they believe that it is necessarily impossible. There are two types of arguments given in favor of this view; one rooted in evolutionary biology and one slightly more philosophical. I'll start with the former.

A somewhat popular folk view of evolution going along with the pseudoscientific interpretation of the social Darwinist motto "survival of the fittest", is the view made famous – perhaps unintentionally – by Richard Dawkins as the "selfish gene" view. According to Dawkins, in some indirect sense, "we are born selfish" (2006, p.3). Another famous quote reminiscent to this view is Michael Ghiselin's "Scratch an altruist and watch a hypocrite bleed." (1974, p. 247). There are two big problems with this view. There are obvious problems with talking if as genes could be selfish, because it requires them to have a *self*, which in turn presupposes that genes have interests. Genes do have the tendency to replicate, and that could be said to allow for talking as if the gene has purpose. The road then becomes short for those who want to take a leap and talk as if the gene has an interest in replicating. However, this doesn't mean that the view that genes have interests are justified. Even if we allow the metaphor, inferring that "we are born selfish" still need to explain why we should conflate the interest of a person's genes with *her* interests. In advancing the interest of my friend's genes, I wouldn't necessarily be helping *her*. As Joyce (2006, p. 15) writes "*…to confuse a person with her genes is as silly as confusing her with her lungs or her lymph nodes*". More precisely, it confuses the *cause* of a mental state with its *contents*. It's a common mistake amongst some evolutionary psychologists who think that evolutionary explanations uncover the *bona fide* content of our desires, motivating reasons and interests. If my fearlessness about speaking to a large crowd is partially caused by the three shots of tequila I just had, it would be silly to draw the conclusion that I am fearless of the tequila. Yet this is what's being done when one tries to explain a person's interests in terms of the interest of her genes. Perhaps the mistake stems from an ambiguity within the notion of 'a reason'. When my children are sick, it's their suffering that figure in my deliberations that ultimately constitute *my reason* for caring for them. It might be that *a reason* for why their suffering motivates me is due to the fact that caring for my offspring is fitness increasing and has been selected for in the species that I am a member. It does not follow, however, that this was therefore *my reasons* – the considerations I acted upon. These are two different explanatory levels and should be kept apart.[6]

A stronger argument against the possibility of psychological altruism is a conceptual claim about intentions. The proponents of this view are persuaded by the simple argument that even when one is acting intentionally one is ultimately acting on egoistic grounds. If one is acting intentionally one is identifying a desired outcome and striving to satisfy you own desires is

[6] I am aware that the cause/reason dichotomy does not fit well in this context, but for the sake of argument I hope that the reader can forgive my usage here.

selfish – that is what it means to act intentionally. On the surface, the argument has traction. However, it is quite possible that we have *different* types of desires. Some can be directed towards ourselves and our own wellbeing, while other types of desires can be directed towards others. What kind of process generates the intent of a desire or goal? Consider a scenario S where you are in a room alone, hungry and there is a basket of bread on the table. In this scenario you want to eat all of it. Consider then a counterpart scenario S* where there is another hungry person in the room. In S* you respond to the perception of the needs and wants of someone else and want to share half the food with the hungry person. Are you genuinely altruistic or are you calculating that *behaving* altruistic will lead to some desired self-interested future outcome? If you are a behavioral altruist in this example, your desire in S and S* are unchanged: your desires were self-directed in both. If you are a genuine altruist, then your desire in S is different from that in S*: your desire will have shifted from self-directed to other-directed. The possibility of altruism, then, depends on what happens when you respond to the perception of the needs and wants of someone else. For the egoist case to be credible, the response consists in a Machiavellian-like calculation. For the altruist case to be credible, there need to be an emotional response partly shifting the psychological state of the altruist.[7]

Without going deeply in to the character of the emotions (the role of the emotions in moral judgment will be discussed later in many parts of the thesis), I think we have some reason to believe that some kinds of emotional response can be understood as basic altruistic emotional reactions. A change in emotional states respond to a change in our physiology, for example via hormonal regulation. When someone reacts with resentment to some insensitive comments, that person will undergo physiological changes. The specific causal details of such connections are, of course, at this point, matters of speculation, but even before knowing them we can *"...reject an approach to the emotions that would leave out either the physiological or (...) cognitive features"* (Kitcher 2011, p. 27). It has been argued that certain human emotions and their expressions are universal (Ekman & Robinson 1994). This means that there are several basic reactions, found in all human societies, that give rise to the same facial expressions. According to Paul Ekman and Richard Davidson (1994) these are anger, disgust, fear, surprise, happiness, sadness and contempt.[8] To account for their complexity, Paul Griffiths (1997) name Ekman and

[7] This way of describing an altruistic state is due to Kitcher (2011, ch. 1)

[8] I am aware that Ekman and Robinsons work is not entirely uncontroversial, and that there is contemporary work critical of their views. However, they are useful for the point I am making and their basic claim that we have some basic emotions is plausible.

Robinson's basic emotions 'affect programs'. Each of the basic emotions or affect programs are triggered by environmental factors, which give rise to inner emotional states and prompt expressions by activating perception. The debate to which such basic reactions have a genetic basis is not settled, we could imagine that the environments – physical and social – that would not give rise to such reactions just hasn't obtained. But I will allow that human beings who develop in all known human societies will share dispositions to basic emotional reactions, such as anger and happiness, and they will display similar facial and bodily expressions characteristic of such emotional states. This doesn't mean that the same thing or event will cause disgust all human beings: all feel disgust, but different people find different things disgusting. Neither does it entail that such states lack any cognitive aspects. Even if every emotion involves some basic emotional state, *"…a large diversity of emotions might be distinguished by the cognitions (…) connected to that state"* (Kitcher 2011, p. 28).

Thus 'the perception of the needs and wants of someone else' by a genuine altruist must involve some basic emotional state. There is some independent evidence for this. Consider the phenomenon described by moral psychologist Martin Hoffman (2000), that when new-born infants lying together in a room at the nursery, they react to the crying of others. If one starts to cry, it could set of the whole room. What is the explanation? Suppose that crying signals unhappiness and that makes all the other infants unhappy as well. Or put other terms, the noise picked up by the infant's eardrum connects with some sort of neural mechanism affecting the emotional state of the infant, spreading the misery of one to those around. The science of such neural mechanisms has recently made significant progress. Neurophysiological experiments and brain-imaging (primarily on macaque monkeys) found a variety of visuospatial neurons respond to the action we observe in others. Strikingly, the same neurons fire in the same way when we recreate the action ourselves. It seems that we have a capacity to imitate or mirror[9] the behavior perceived in others (See Acharya & Shukla, 2012 for discussion and references). For example, my observation of your facial expressions produces neurological firings and result in my imitation of you, which causes changes in my basic emotional state and physiology. A psychological altruist has, then, a specific relational structure in their psychology – when in the presence of others, the altruist is disposed to change and align her desires, hopes and intentions more closely to that of the others. Going back to the case for the egoist, it is highly implausible that Machiavellian-like calculation enclosed within the self can account for the complex

[9] The neurological properties have come to be known as *mirror neurons*.

emotional states human beings find themselves in, and their connection with perception and cognition. It simply unrealistic that behind every altruistic behavior, there is some sort of "cold" reasoning or Machiavellian manipulation that override all emotional states. This doesn't mean that altruists give everything, and egoist nothing. The debate is often cast in terms of complete division. The more plausible view is that the change that occurs from S to S* or similar situations, can vary in terms of intensity. The outcome will thus depend on many factors, and the alignment of the altruist's desires to that of the other will come in degrees. The important thing for my account is that altruism is a real possibility.

1.2 The natural selection of cooperative traits

In the last section I hope to have removed some prima facie reasons why we shouldn't think that psychological altruism is a real possibility. I ended with an appeal to how our basic emotional states enables such a capacity. I think we readily can extend this way of reasoning to other 'prosocial emotions' such as love and sympathy etc. as well. This wouldn't do much for my *evolutionary* argument if I cannot show how natural selection, in some circumstances, would have favored such traits over more self-serving traits. In this section, then, I am going to outline some specification on how psychological altruism can be given an adaptive explanation in genetic terms: whether the existence of the trait is to be explained by reference to a genotype that gave our ancestors reproductive advantages[10]. Without the power of natural selection, the case for constructing an evolutionary basis for explaining our capacity to make moral judgments would not be credible. In saying that I am going to *show* how cooperative traits might have been selected for, I would like to make the disclaimer that I do not at any rate reserve the right to call myself an expert on the topics: biology, evolutionary psychology, natural selection, evolutionary theory etc. The following section consists of a relatively quick and short exposition of well-known and popular works in evolutionary theory. For someone studying some of the fields mentioned above, I could imagine that most of the theories I will discuss are typical 'textbook' material.[11] Therefore, my own argumentative tone will be slightly dampened

[10] In saying that altruism was adaptive to our ancestors, I am not claiming that it continues to be adaptive. It is possible that, perhaps like our taste for fatty foods, it is no longer adaptive. (see Dawkins 2006, p. 220-21 for discussion)

[11] My own introductory level of familiarity with the theories concerning the evolution of cooperative traits first came from reading Michael Tomasello's brilliant book *A Natural History of Human Morality*. In the aftermath of reading this book, I benefited greatly from reading Joyce (2006), Kitcher (2011), Rosenberg (2002), James (2010), Boyd and Richerson (1985) and not least Sober and Wilson (1999).

in the paragraphs to come – my job will rather be to cash out some philosophical consequences later.

Although I reflected critically of Dawkins 'selfish gene' view above, my arguments weren't meant to target his main thesis. Dawkins did argue that because the genes are evidently selfish (that replicating themselves is all they are up to), we could readily suppose that this penetrates to the individual level. However, viewing genes as selfish serves another purpose: to locate the *level* at which natural selection works. Dawkins' main objective was to argue that instead of thinking of organisms as the unit on which natural selection works, we should think of them as *vehicles* by which genes succeed in reproducing themselves. Clearly, it is a genetic mutation that allows the cheetah to run faster eventually leading to more cheetahs with more copies of that gene. But what happens when what is good for the genes is bad for the organism? What happens when a cheetah mother refrains from feeding to protect her kittens from a predator? Since her kittens share over 50 percent of her genes, such behavior benefits the survival of her genes, but surely it amounts to *fitness sacrificing* behavior. Some simple Darwinian logic should deem that such individuals, with her genes, is headed for extinction. However, in 1966 evolutionary biologist William Hamilton demonstrated that when there is such a 'conflict of interest' (If I could, only for a moment, be allowed to speak of the genes as having interests) between the individual and its genes, natural selection will tend to favor the genes. Because the kittens share so much of her genes, and since the genes responsible for her fitness sacrificing actions toward her kin are likely to be among those genes, ensuring the survival of her kittens is a way of replicating her own genes – including those that are responsible for the fitness sacrificing behavior. Therefore, what is bad for the organism in terms of fitness, might be good for its genes. Hamilton went on to show with mathematical rigor how natural selection might favor actions that are fitness sacrificing toward kin. Such an action might be favored by natural selection if

$$rB > C,$$

where r is the degree of genetic relatedness to the individual, B is the benefit to the recipient and C is the cost to the individual (Hamilton 1964). This evolutionary mechanism, called *kin selection*, is plausible and confirmed theory about how and why fitness sacrificing behavior toward relatives might have been selected for. However, it can't by itself explain such behavior toward non-kin. If the r in Hamilton's rule stated above is zero, then the same will follow for

rB and automatically the action is too costly to perform. Therefore, it can only partially explain altruistic behavior – we are left with explaining how fitness sacrificing behavior can extend to non-kin. As we shall see soon, kin selection acts as a basis for other theories that might to this job. But before turning to those, there are more direct ways that kin selection may be an important factor in explanations of cooperative behavior toward non-kin. Notice that to be able to help one's kin, a creature must be able to recognize who one is actually related to. Some species recognize kin by scent, but it's safe to say that human beings don't normally recognize kin in this way. One hypothesis is that kin detection is not innately specified, but that kin is presumed to be those conspecifics an individual most regularly interacts with – which is very plausible. Such that the kin selection mechanism is the capacity for fitness sacrificing behavior toward those one most regularly interacts with. If this is true, given the societies we live in today, with tightly linked relationships with non-relatives from early childhood, we can see how the mechanism might be used for fitness sacrificing behavior toward non-kin.

Yet such an explanation doesn't explain how human beings, or even other animals, engage in altruistic behavior with conspecifics which is non-kin, or someone that one does not interact with frequently. The kin selection mechanism might explain why a monkey might spend the whole afternoon grooming a family member, but it has problems accounting for why we see monkeys doing the same for non-kin. Grooming non-kin may give the beneficiary more benefit than the cost taken by the groomer, but there is cost nonetheless. What could possibly be the advantage for the cost taker? The answer is *long-term cooperation*. Taking the cost of grooming non-kin for an afternoon is a small price to pay for the benefit of being a member of a larger reciprocal scheme. Say that you use one day of the week grooming others, but in return you would get groomed by others the remaining six days. Potentially, the benefit received massively outweighs the cost. Engaging in what might initially appear to be fitness sacrificing behavior, then, actually might be fitness advancing. However, this form of cooperative behavior does presuppose that an individual must act on the basis of an expectation that the fitness sacrificing behavior will be reciprocated. If not, how could such schemes possibly evolve? It's easier to understand that when such a scheme already is firmly in place, new members can more vividly see and recognize its reciprocal nature and the benefits it bears. It's more difficult to see how the scheme might get going in the first place.

Basically, the situation is very much like that of the famous game-theoretic scenario *the prisoners dilemma*. The classical story presents a situation in which two burglars are captured

near the scene of a burglary and are interrogated separately by the police. Each must choose whether or not to confess and blame the other. If neither confesses, and therefore cooperates with the other, both will serve one year on charge of carrying a concealed weapon. If both confess and blame each other, they both serve ten years in prison. However, if one of them confess and blames the other, while the other burglar does not confess, the one who has collaborated with the police is free of charge, while the other must serve twenty years in prison. It's well known that defecting is the default strategy because one cannot assume that the opponent is going to choose cooperation. The problem of assuming that your opponent in the dilemma will act in the same way as you are perhaps not surprising given that the game is also rigged such that you have no previous experience in the game to draw on. Additionally, the fact that the consequences are so final might also affect what choice each player makes – if you knew that there would be second chances, would you then opt for a different strategy?

Perhaps the one-shot nature of the game explains why in that particular situation noncooperation is the best strategy, but this might not be very helpful in explaining the evolution of cooperation in general. If the prisoner's dilemma would hold this explanatory power, it should rather be played in repeated number of times in relatively stable and small social groups because these are the conditions under which human beings developed. What would happen if each player faced every other player a certain number of times? In this iterated game, several strategies are available. One could, as in the original prisoner's dilemma, either cooperate or never cooperate. Or one could "test" cooperation by cooperation in the first game with any player and then in the next game do whatever she did the last time – a strategy that came to be known as "tit-for-tat". In the computer simulations famously conducted by the political scientist Robert Axelrod (1984), the optimal strategy for most iterated prisoner's dilemma games, the strategy that provides the largest total payoffs, is the "tit-for-tat"-strategy. If one adds the feature of eliminating the least successful strategy after a certain number of rounds, the tit-for-tat strategy yields even greater payoffs. This is consistent with the assumption that the evolution of human cooperation most likely is made up of extensive trial and error, where we dispose of less successful strategies as we experience the consequences of utilizing them. Among humans the "tit-for-tat" strategy is effective for several reasons. First, it doesn't require a lot of cognitive skills to identify the strategy of the opponent. Second, the fact that it begins with cooperation makes it a soft and nice start. Third, its provokable – it responds to defection by defection on the next round. Fourth, its forgiving – once the partner has cooperated again, it returns to cooperating.

It appears that "tit-for-fat", first and foremost a *game-theoretic* strategy, is also an *evolutionarily stable strategy*. The group of players playing "tit-for-tat" for themselves won't be vulnerable to excessive free-riding invaders who never cooperate. The beauty of "tit-for-tat" is that even if free-riders win the first round, they will lose each subsequent round in the game. Thus, free-riders will always ultimately fail to get a foothold, and in the long run they will be eliminated. Once the "tit-for-tat" strategy gets enough footing in a group of individuals, it will inevitably spread until it becomes the dominant strategy, and it will eliminate all other strategies. The explanatory power of "tit-for-tat", among other reasons, lead evolutionary biologist Robert Trivers (1971) to suspect that there might be a genetic basis for our dispositions to engage in reciprocal cooperation. The fact that less successful strategies are eventually eliminated might correspond to those who lacked the genetic dispositions to reciprocate and who were thereby less reproductively fit. Trivers suggested that the supremacy of "tit-for-tat" corresponds to the spread of individuals with reciprocating genes throughout a population. One might object that the idea that tit-for-tat, or something like it is built into our genes, might over-reach because it doesn't seem to sufficiently account for our fallibilistic nature. Players might misperceive the other player and think that she is defecting, or one might just defect by mistake. However, given that such reciprocal scheme is potentially very beneficial for increasing fitness and that we now have an explanation for why the scheme might get started, it's very likely that we are have innate mechanisms for altruistic behavior with the expectation of reciprocation.

More pressing is the weakness of the "tit-for-tat" strategy to account for larger-scale cooperation in large groups. It does well in relatively small and stable groups made up of purely self-interested individuals. However, in larger groups with several selfish individuals, the strategy does much poorer and the selfish individuals can easier end cooperation. Still we don't have an explanation of how more indiscriminate altruism we observe in contemporary societies can develop. With kin selection and reciprocity with have an account of why the relevant genotypes might be spread by increasing the fitness of cooperators, but these are not strictly theories about the psychological motivations of individuals. As it stands, all they amount to are *biological altruism* and not *psychological altruism*. We need to supply a plausible story in which these traits can amplify and be sustained at group level. This is exactly the kind of story provided by philosopher Elliot Sober and evolutionary biologist David Sloan Wilson in their theory of *group selection*. Everyone agrees that there has been a

lot of kin selection and reciprocity throughout the history of life, but its more troublesome to imagine scenarios were more self-serving traits don't deteriorate cooperation at large group-level. It's very implausible that natural selection would favor a genuine fitness sacrificing creature of a fitness advancing creature. However, Sober and Wilson (1998) argue that, say in terms of kin selection, it is plausible that that biological altruism grounded in inclusive fitness translates to psychological dispositions to care for kin. That such psychological capacities are adaptive could be demonstrated by looking at kin selection through the prisoner's dilemma. Philosopher of Biology Alex Rosenberg (2002, p. 324) summarizes the point neatly:

> *"In a one-shot prisoner's dilemma involving kin, therefore, both may be advantaged by cooperation regardless of the other's action, if the pay-off they are 'designed' to maximise (reproductive fitness) satisfies the inequality r > b/c, where r is the coefficient of relatedness (1/2 in the case of offspring and siblings, 1 in the case of identical twins, 1/4 in the case of cousins and nephews), b is the pay-off to mutual cooperation, and c is the cost of cooperation in the face of selfishness. If the group's fitness is a function of individual fitnesses, then groups of kin-related agents playing the cooperative (or 'sucker's') strategy in a one-shot prisoner's dilemma will be fitter than groups composed of pairs of mutual free-riders playing the defector strategy, and also fitter than mixed groups of pairs of free-riders and suckers"*

Sober and Wilson (1998) argue that the conditions under which kin-related agents were able to pursue such a strategy has obtained in hominid and human history. A very rough and simple sketch of such a story is that it all started with very small packs of kin-related individuals which over time deployed the cooperative strategy to aggregate into larger fitter groups. Once the individuals can recognize kinship, and not least other cooperative individuals, they have sufficient cognitive resources to seek and find each other, forming successful groups were cooperation is the norm. The group can later introduce secondary reinforcement behaviors to weed out potential freerides. Such enforcement behavior can be shaming, confiscation and reporting. The crucial contribution of their theory is to show how conforming to the norms of enforcement is far less costly to the enforcers than breaking the norms of cooperation, creating strong selective pressure on cooperative groups. The long-term result is a correlated equilibrium of groups cooperators having driven the groups of defectors and free-riders to extinction. Sober and Wilson demonstrate this pathway with statistical rigor in their work.

The conclusion to draw from this is that selection might not always work on the genetic or individual level, but it might also work at the group level. In *The Descent of Man*, Darwin himself hinted at group selection. He pointed out that tribes with members willing to sacrifice themselves will outcompete tribes without such self-sacrificing members. However, he recognized that even in altruistic tribes there will always be members who refuse to sacrifice. If this is so, such individuals seem more likely to survive and have children. Over time, the heirs of the selfish individuals will after generations of natural selection surge and overrun the altruistic group. Like Darwin, Sober and Wilson admits that group selection is only one among many likely ways that natural selection works. What they show is the conditions under which group selection is a possibility: they argue that as long as an individual's costs is outweighed by benefits to his group, altruism can evolve. They ask us to consider the leaf-cutting ant *Acromyrmex versicolor* as a prime case of when an individual's cost is offset by benefits to his groups. The leaf-cutting ant forms different colonies with several females in each colony. For the colony to thrive it needs leaves to grow fungus to feed its offspring. The crucial observation is that in each colony one of the females is given the task of gathering the necessary leaves. This lonesome search for leaves involves high degrees of risk and the forager will much more likely be eaten before she can have offspring herself. However, it is good for the colony. According to Sober and Wilson, this is an example of when group selection cancels out individual selection. The fact that the females aren't related means that it cannot be explained by kin selection. That doesn't mean that females do not compete with each other in other ways, Sober and Wilson recognizes competition between individuals. It means that the unit of which natural selection operates might come in different levels. There are models that suggest that ordinary group selection too is insufficient for the kind of large-scale cooperative groups we observe in human society. Those critics claim group selection is involved but that other factors such as sexuality or culture (possibly amongst other factors) must figure in our models to account for the indiscriminate forms of altruism we are familiar with. However, what I have said suffices to show that its very plausible that our capacity for psychological altruism was selected for by biological natural selection. I will return to the question of how group selection might be combined with other factors in section 2.4.

1.3 Taking stock: Distinguishing "wants" from "oughts"

I don't mean to give the impression that the mechanisms I have outlined above is in contention for *the* evolutionary process responsible for our evolved cooperative traits and specifically our capacity for (psychological) altruism. For all we know, it could be a mix of all such process plus many of which remain untheorized, or it might just be one of them. Different cooperative traits could be due to different evolutionary mechanisms and some might even be built on top of each other. Because knowing the *source* of a pressure for cooperation (say kin selection), doesn't tell us exactly *how* the behavioral trait (altruistic behavior) might be achieved. For one thing is certain: natural selection cannot bring about such behavior magically out of thin air; it goes to work on whatever mechanisms that exists governing the organism's behavior, tweaking or transforming them to boost stronger or create new behaviors (Joyce 2006, p. 44). This "messy" picture of evolution could also help us see why adaptive psychological mechanisms can sometimes involve specifically targeted conditional motivations, for example discriminate altruism towards kin or reciprocators, while in other cases the most cost-effective way to promote adaptive cooperative behavior in an environment is *less* discriminate and perhaps even *unconditional* forms of altruism (See Kitcher 1998, 2005). In circumstances where it is cognitively hard to recognize the relevant attributes (kin, group membership, reciprocators etc.) the most effective way of promoting fitness might be an indiscriminate form of cooperative behavior. This hypothesis can also address the concern that it is hard to make sense of some contemporary cases of altruism in evolutionary terms. I am thinking about the evident altruistic behavior towards complete strangers; for example, when we donate money to aid organizations to help people on the other side globe who are in no position to reciprocate. This might strike one as mysterious from a biological point of view, because such indiscriminate altruism isn't adaptive in the way it might be for more conditional altruism might be; helping children on the other side of the globe isn't going to multiply your genes. However, a trait that isn't currently adaptive might have been so once. Even in the local environments that our ancestors lived in a relatively indiscriminate form of psychological altruism might be to the most benefit of one's kin or potential reciprocators, and so could altogether be the most effective adaptive mechanism. An adaption that is no longer adaptive isn't a strange phenomenon at all; our taste for fatty or sugary foods are familiar examples of traits that isn't no longer beneficial in our contemporary world (Dawkins 2006, p. 220-221).

If human reproductive fitness was enhancing by a proclivity for helping others, even complete strangers, what might the process of natural selection have done to our brains in order to achieve this? The answer I have been building up to is this: psychological altruism. Coupled with the

fact that psychological altruism is importantly linked with the emotions, we thus have an evolutionary explanation for why we might have neural mechanisms that provide the motivation for 'responding to the perceptions of the needs or wants of someone else'. This is only the first step toward solving the task of explaining how evolution might have something to do with our capacity for moral judgment. To see this let me conclude the altruism theme for now by making a crucial distinction between doing something because you *want to* and doing something because judge that you *ought to*. Just because we have identified a potential *inhibition* against non-cooperation such as stealing or killing etc. doesn't mean that we then have secured a notion of *prohibition*: the idea that we shouldn't do those things because they are wrong (Joyce 2006, p. 50-51). To judge that abortion is wrong is not to express that one does not want to have an abortion, rather, it is to assert that abortion is wrong and shouldn't be done in general. A related distinction is that between what is *accepted* – which depends on the response of one's peers – and what is *acceptable* – what one judges to right or wrong *independently* of one's peers. This doesn't mean that psychological altruism and other prosocial emotion are irrelevant to moral judgments: an altruistic act will indeed often involve a moral judgment. It only means that the moral judgment is not a necessary part of it. This distinction will emerge in the next chapter when discussing the issue of nativism, and the topic of the emotions will follow us throughout the thesis. The important conclusion from this theme that will follow us onwards is that we have an evolutionary explanation for why human beings have come to have prosocial inclinations and aversions and that our capacity for moral judgment needs further explanation.

1.4 Non-cognitivism and cognitivism about moral judgments

Any attempt to understand the evolution of morality would suffer from ignorance would it lack an understanding of what morality is. There are different senses of what we understand by "morality". For example, we could mean being morally praiseworthy, but our focus here will be on the moral *judgment* – the capacity to judge in moral terms in general. In answering the question of what a moral judgment is, I will not at all attempt to give a complete conceptual account of a moral judgment – that is, supplying all its necessary and sufficient features that would pick out moral judgments in any possible world. Many larger piles of paper than this have been devoted to this task. My strategy will rather be to identify some its crucial features and use this as my basis throughout the thesis.

The first thing to note is that the issue of moral judgment is extremely complex and there are vast amounts of views on the matter. Scholars struggle even to agree on what kind of a thing a moral judgment is. Some hold that it is a type of mental event while others insist that it's a sort of linguistic utterance. If we go deeper and focus on (say) the camp that thinks that the moral judgment is a linguistic utterance further disagreement follows. What type of utterance are we talking about? Some claim that moral judgments displays our feelings, others that they express commands. Others say that moral judgments report facts, but which facts exactly? On one side answers varies from facts about the speaker or her culture, or facts about God's commands. While another side claim that moral judgments report objective mind-independent facts. But then again there is the question of what kind of mind-independent facts? Are they physical, epiphenomenal or maybe *moral*? There are many alternatives. Desperate to avoid a metaethical quagmire, I will dodge these controversies by accepting that a moral judgment can be both a mental event and a linguistic utterance. But there are some theoretical options that I wish to exclude. As a first approximation to answering the question of what a moral judgment is, I want to approach the moral judgment as a kind of speech act or public utterance. The rest of the thesis will involve the topic of moral psychology. I want to allow that moral judgments as a public utterance express both cognitive and conative states. Thus, I want to exclude two options located at each extreme in the debate between cognitivism and non-cognitivism. This means rejecting both theoretical options in their pure form.

Non-cognitivism is the view that moral judgments as a kind of utterance do not express beliefs but performs some other kind of speech act. As a semantic thesis, non-cognitivism holds that moral predicates like "is wrong" or "is right" is nothing but grammar. The predicate disappears when we get under the surface and find the "real" meaning of a moral judgment. According to the pure form of non-cognitivism the moral judgment "killing people is wrong" does not express a proposition that could be true or false, but rather it is as if one were to say "killing innocent people!!" with a tone of voice indicating a special feeling of disapproval was expressed.[12] I think that there is something right with non-cognitivism – moral judgments do involve our emotional states in very important ways (which will be discussed frequently thought the rest of this thesis). However, in its pure form, some important features are lost. Consider, for example, what it means to describe someone as a "Yankee" contrary to describing someone

[12] Famous non-cognitivist A. J. Ayer could be said to be a proponent of a pure non-cognitivism.

as “American”. To say “Ray is an American” is to describe Ray as having a certain nationality. To say “Ray is a Yankee” is both to describe Ray as having that heritage and to express a derogatory attitude. Asserting “Ray is a Yankee” is not the same as merely saying “Ray!!” in a clear contemptuous tone because we are both expressing the belief that Ray is American and contempt not just for Ray but all Americans. The pure non-cognitivist interpretation of the predicate “is a Yankee” is implausible. This should lead us to suspect that the pure non-cognitivist interpretation might be implausible for moral predicates for the same reasons. When we are saying something like “killing innocent people is wrong” we are both asserting something about killing innocent people and expressing a conative approval for a standard that condemns it.

Before I argue further for the view that a moral judgment can express both cognitive and conative mental states, we should understand properly what we mean when referring to an *expression-relation*. Sometimes we might be using the word ‘express’ to denote a causal relation. We may say that in slapping Ray in the face, Nora expressed her anger, and we may mean that anger caused her action to do so. Sometimes, however, ‘expressed’ is used differently. Later on, Nora might apologize for slapping Ray and in doing so she expresses regret. Even though the apology is insincere, she might succeed with her apology. She expressed regret either way, because an apology *is* an expression of regret. This way of using “express” does not refer to a causal relation between Nora and her mental states, but rather denotes a complicated circumstance with relations between Nora, Ray and a bunch of linguistic conventions.

With this understanding of the expressing-relation, we can say something about when we should expect a sentence to express a mental state. If a sentence is uttered in some circumstances where it functions to express a mental state and the speaker immediately and explicitly adds that she doesn’t have that mental state, we should expect instant confusion in her audience. If no further explanation of the confusion is imminent, then we have few options but to take it as evidence that the statement functions to express that mental state. Consider the following statement in circumstances where “Sorry” alone is an instance of a successful act of apology.

(1) Sorry. But I don’t regret it.

This statement is odd in different ways, but for our purposes (1) is strange in a special way. The first sentence “Sorry” *expresses* a mental state and the second sentence “But I don’t regret it”

reports a mental state. But the most peculiar thing to note is that the sentences disagree. The statement has a flavor of what has been come to be known as *Moore's paradox*, after G. E. Moore and his introduction of statements like "The cat is on the mat, but I don't really believe it". The paradox consists in the fact that such sentences involve no contradiction, but the second sentence in the pair seems to nullify the first, leaving the listener confused as to what should be assumed about the speaker's attitude. The next statement is, I think, an instance of the same phenomenon:

(2) Killing innocent people is wrong. But I don't believe that killing innocent people is wrong.

Someone who uttered (2) would I suspect evoke much the same response as someone who uttered (1) or more generally any instances of Moore's paradox, given that if they had not added the second sentence they would naturally be interpreted as making a moral judgment. The point against pure non-cognitivism I am trying to make is that the confusion arises because we naturally interpret the first sentence of either (1) or (2) as an assertion – that is, an expression of belief.

This doesn't mean that we should swing the pendulum too far to the cognitivist side of things, for its equally implausible that moral judgments always are mere assertions. Consider this sentence pair:

(3) Killing innocent people is wrong. But I subscribe to no moral standard that condemn such actions.

This sentence pair is equally as odd as (1) and (2). It seems to me that they have a Moore's paradox-type peculiarity, which means that moral judgments are not merely assertions, but they express both beliefs and conative states. One might respond that (1), (2) and (3) are being uttered in circumstances where the speaker is a joking or playful manner like a sarcastic tone of voice, therefore there is no real paradox in play. This is not a successful counterexample because even though there might be a convention according to which a Moorean paradox doesn't incite confusion because it's a joke, that's only a one-time achievement. It doesn't follow that there never are circumstances where the paradox results in confusion. For example, even though the word "rascal" functions as a pejorative, one could swiftly shift to a playful

context by adding a blink of an eye following the utterance of "you rascal" and neutralize all offensiveness. However, this doesn't undermine the fact that there is a linguistic convention according to which "rascal" expresses contempt.

The view I want to hold for the rest of the thesis is therefore that moral judgments express both cognitive (beliefs) and conative (non-belief) states. By this I am rejecting a pure form of non-cognitivism – according to which to say that someone is evil is just expressing a feeling like "this person booo!!". Secondly, I am rejecting a pure form of cognitivism, according to which saying that someone is evil is just ascribing that person with a property, like being tall or being American, and not thereby expressing any negative feelings about that person. Moreover, saying that someone is a bad person is not at all like making a neutral statement like saying that someone left the restaurant without paying. The latter statement leaves out any attitude toward the culprit. On the contrary, if the former statement is put forward without any ironic facial expression etc. we know that the speaker disapproves of the person in question.

One might worry that this dual view of moral judgment leads to the unattractive conclusion that any judgments express a mental state with both belief-like and desire-like aspects – a "besire" to use J. E. J. Altham (1986) famous term. But since we are holding that the expressing-relation is non-causal we avoid any dubious modal relations between the speakers cognitive and affective states. Therefore, to clarify, the claim is not that the cognitive/non-cognitive duality of moral judgments express a single state, but rather that it expresses *two* mental states. I am not at all claiming that this is an attempt at a final answer in the debate between cognitivism and non-cognitivism. However, the conclusion is an important one and will follow us as a premise throughout this thesis. For now, we can note that any successful strategy for an evolutionary vindication of morality will have to account both for its cognitive and non-cognitive aspects.

1.5 The practical authority of moral judgments

In the last section we considered some implications for moral judgments by thinking of it as a linguistic convention. However, it is normally thought that morality transcends arbitrary conventions. As Joyce (2006, p. 57) himself puts it

> "*...there is surely more to the practicality of moral judgments than just a linguistic arrangement whereby the choice of one predicate over another will allow us to express our attitudes as well as communicating what we believe*".

In this section we are going to consider a simple and perhaps uncontroversial observation that morality appear to have a kind of special status (whether morality deserves this special glow is a different matter entirely and one that I will discuss later). The special status I am referring to is the apparent *practical authority* of moral judgments. Calling something *morally* right or wrong isn't something that we normally can just ignore or evade in our practical deliberations and interactions. I will argue that the practical authority that morality purport to have is perhaps its most unique and noteworthy feature and plays an important role in distinguishing it from other human conventions. However, it could be that as a matter fact morality is just a matter of human convention and that the practical authority morality appear to have cannot be cashed out. This is a problem we will discuss in later chapters. What I want to argue here is that we normally think of moral judgments as transcending human conventions – that is, when we are trying to understand the basic features of moral judgment we are acknowledging the presence of practical authority.

When we are expressing a moral judgment, we are putting forward a deliberative consideration. Suppose that I am a passionate anti-abortion protestor. As I am walking past a group of pro-abortion activists and asserting to them "Abortion is morally wrong! Your actions are evil!", I am aiming to say something that demands consideration irrespective of the activist attitudes toward me. Of course, we could readily imagine the pro-abortion activists not being moved at all by the moral judgment I just expressed – they think that I have said something false. However, even if they disagree with me they would have to acknowledge that I have put forward a *moral judgment*. By acknowledging this, they are also recognizing that I am introducing a practical consideration of importance. By expressing my moral judgment, I am demanding that it be relevant to other's deliberation. That is another reason why moral judgments can't be just expressing our feelings (like is supposed by pure non-cognitivism). Unless one cares deeply about someone else's inner states, why should the expression of one's conative attitudes be relevant to one's deliberations? Instead of expressing the moral judgment I did before, let's imagine that I rather *reported* my disapproval by stating "Your actions spark a feeling of disapproval in me" or that I *expressed* my disapproval by saying "Abortion boo!!" Could we blame the anti-abortion activists to respond simply "Ok, but so what?" I think not.

Simply reporting or expressing the fact that some feeling is being aroused in you is never in itself to provide a practical consideration. It would be odd to think that it would provide the anti-abortion activists with a reason to stop their activism as the moral judgment would have done.

Where does this *practical clout* come from, to use Joyce's apt technical term? It's widely regarded that it has something to do with the apparent *inescapability* of moral prescriptions. By this I mean the proper scope of application of a moral demand can extend to an agent regardless whether that agent has any existing inclinations or desires that would be served by conforming to that demand. To judge that Norwegian officials has violated morality by jailing foreign nationals without charges, is not necessarily to refer to some inclination or desire those officials have that would be satisfied by doing their moral duty. Even if we assume that they have no such inclinations or desires at all, we may still judge so. Our moral language acts as if there is a moral duty that applies regardless of whether there are any such existing motivations. At this point we way legitimately ask where such inescapability comes from? What is its source? This is a notoriously difficult question and controversy lies at the heart of any substantial answer. One important strand of answers emphasizes the role of sanction to determine the applicability of moral judgments. For example, it is widely regarded that the concept of punishment has an important role to play in the evolution of the moral judgment, something that will be discussed later in this chapter. Generally, we think that a moral transgression deserves punishment. However, as Joyce (2006, p. 59) correctly points out, that moral transgressions deserves punishment must not be confused with the thesis that the applicability of moral judgments depends on *"…the unpleasantness of punishment, or on the pleasantness of reward*". If moral judgements contained prescriptions for avoiding punishment or for reaping rewards it would be like practical advice, where its applicability depends on there being an desire on behalf of its recipient that would be satisfied by the advice. That isn't to say that something can't be both morally *and* prudentially wrong at the same time, but this fact shouldn't overshadow the crucial distinction here. Which is that when we say that harming innocent people is morally wrong, we include people who couldn't care less about suffering the consequences of harming innocent people.

A widely influential theoretical move that has been but forward considering the recognition that moral judgments are typically not practical advice is to think of them as categorical imperatives, as opposed to hypothetical imperatives – a distinction from the legacy of Kant. For Kant, a

categorical imperative is a practical imperative, but one that doesn't derive its legitimacy from some end desired by the subject of the prescription. A hypothetical imperative, by contrast, does depend for its legitimacy on some goal aimed at by the person in question. For example, the imperative "you should eat something" tacitly implies "if you are hungry" or some other reason for why your body needs nutrition. If the recipient lacks a desire for not being hungry, then the imperative will typically be withdrawn. This is a hypothetical imperative. By contrast, the moral judgment "don't harm innocent people" doesn't appear to depend on the desires or inclinations of the recipient. She cannot dodge the moral imperative by introducing deviant desires and ends. Even if she really enjoys harming innocent people others would not accept this at face-value and permit such behavior only because she takes pleasure in it. This is what makes up the apparent inescapabilty of moral imperatives. Even though not all moral judgments are categorical imperatives, it seems to me that they capture something about morality that makes it distinct from other types of evaluative judgments, and that stands in need of explanation.

Kant himself inferred from the apparent inescapability of moral judgments that the moral law is universally and necessarily binding law of practical reason. Therefore, regardless of one's desires and inclinations it is simply irrational to disregard one's moral duty. I am one of those who think that this is somewhat of an overreaction having been persuaded by a simple argument by Philippa Foot. In her 1972 essay "Morality as a System of Hypothetical Imperatives", Foot argues that it's true that moral demands are nonhypothetical in that they apply to people even if they don't have any desires or inclination that would be satisfied by being moral.[13] However, this is not only true for moral demands. The demands of *etiquette* seem to be nonhypothetical in the same way. For example, even if none of your interests are being served by behaving politely, the demands of polite behavior would still apply to you. From the standpoint of the institution of etiquette, your impolite behavior could still be subject to criticism. Thus, irrespective of your feelings and attitudes the demands of etiquette would still apply to you. Some have read Foot as claiming that the rules of etiquette must be followed regardless of the circumstances – just like Kant holds for our moral duties. Foot's point isn't that there are never any reasons that overrides the rules of etiquette. Surely there are circumstances where it would

[13] Foot later developed a different position on the scope of moral demands. See her (1995) "Does Moral Subjectivism Rest on a Mistake?". I admit that ever since Foots article a lot of water flown under the bridge, but I think her early position is very useful in the dialectic I am building up to. Additionally, Joyce also contrasts his own position to that of Foot. Her description of etiquette will figure in arguments later in the thesis.

be right to break the rules of etiquette for the greater good. For example, breaking the rule of not speaking with your mouth full to warn your table partner that she is about to eat a strand of hair. Although you are justified in doing so, you are still *breaking the rules of etiquette* which show that they were still there irrespective of your desires. Therefore, in terms of scope and conditions for application, the demands of etiquette are imbued with the same inescapabilty as moral demands. But it would be a gross overreaction to infer from this that the demands of etiquette are part of a binding law of practical reason.

Foot is hesitant to call the rules of etiquette categorical imperatives because she doesn't think that they have the practical 'oomph' that we normally attribute to them. I am confident to make the empirical claim that it matters more to others that one should be moral than that one should be polite. The problem of the practical authority of moral judgments remain, but we are now able to separate the wheat from the chaff with terminological help from David Brink (1997). Let's say that in addition to having moral *inescapability*, moral judgments also enjoy moral *authority*. For example, for Kant such authority is cashed out as rationality meaning that it would be irrational not to be moral. Let's say that judgments implying moral normativity, then, is judgments that have both inescapability and moral authority.

In these paragraphs, I am barely touching on a very complex, controversial yet crucial topic in moral philosophy. I have tried to not make any substantial claim about the practical authority of moral judgments, only the simple observation that moral judgments have this feature and that need to be accounted for. Backing this claim is a large body of empirical evidence that ordinary human beings identify a clear distinction between what is required from them by the moral as opposed to the conventional and prudential (See Turiel 1983)[14]. In this, Joyce and I agree. However, Joyce thinks that the practical authority is only a superficial appearance and not genuine. It is an adaptive illusion through the courtesy of natural selection. Therefore, philosophically speaking, that authority is not justified. We will return to discuss his views later, but for now I want to assert two premises from this section to follow us throughout the thesis. Namely that i) one of the things that makes moral judgments "special" is the moral normativity they are imbued with and ii) that any evolutionary vindication of morality must show how this normativity is genuine.

[14] The moral/conventional distinction will be discussed in detail in section 2.3.

1.6 The morality/proto-morality distinction: why non-human animals can't make moral judgments

So far, I have put forward some crucial distinctions between inhibitions and prohibitions, the accepted and the acceptable, the conventional and the moral and the prudential and the moral and so on. They are all, I think, getting at what we might call the difference between morality and proto-morality. What I mean with this distinction will hopefully become clearer in the following, but for now I we can say that proto-morality contains some necessary features of morality, but it's not sufficiently *morality*. It involves some of the 'building blocks of morality', but it lacks some as well. In this section I will look closer at this distinction.

I started this chapter with the (empirical) claim that morality largely involves interpersonal relations. More specifically, it seems as if morality is in place to serve community and facilitate human cooperation. However, what are the mechanisms that we use to ensure the social order? Sober and Wilson (1998) emphasize that to most effectively weed out defectors and free-riders, the group of reciprocators will evolve enforcement strategies such as punishment. It's not surprising then, that we find punishment playing a central role in reinforcing morality too. Indeed, what is a moral prohibition if not something that *deserves* punishment of some sort? Without the notion of *desert*, it seems that the notion of a prohibition doesn't make sense. What's the point of deeming something prohibited but not punishable? Or that something is praiseworthy but not rewardable? It seems that incorporated into the relevant sides of the distinctions presented in the starting lines of this section resides a notion of *justice* – that certain behaviors *deserve* certain responses. To see how desert plays a central role in morality, imagine a group of creatures who act like human beings in the ways we have supposed so far in this chapter. They go around employing categorical imperatives to each other like "don't harm innocent people" and "don't lie" even though the recipient of these imperatives doesn't seem to care. They demand that the imperatives nonetheless have practical considerations regardless of the recipient have the right desires or inclinations. However, if someone of these creatures fails to conform to those prescriptions, imagine that, unlike human beings, they don't subject her to criticism. They do criticize others for all kinds of prudential stuff, like forgetting to put on a rain coat when its pouring or forgetting to eat before a long hike. But they don't see that the criticism that human beings normally subject to moral wrongdoers as something that the situation demands. The point is that these creatures are not like human beings, and I would

argue are not *moral* creatures[15]. Imagine if these creatures were to publicly humiliate another for reasons of deterrence even if the victim were believed to be innocent. Let's say that we were to complain that this was unjust, and she didn't deserve to be treated that way. They simply wouldn't know what we meant, because they lack the necessary conceptual resources for grasping our complaint.

Believing that an action *merits* punishment is not the same as believing that an action will *provoke* punishment. The latter only requires someone to believe that an action will normally be followed by another type of response – namely, punishment. This is merely the perception of regularity in just the same way as we recognize regularity in many ways. For example, that the sun going up will be followed by the sun going down and so on. But it would be silly to think that the sun going up justifies the sun going down. The point is that some creatures might recognize social regularities – that some action is followed by another (say punishment) – but fail to recognize that punishment is justified. Thus, even if you *don't* believe that abortion is morally prohibited, you could still anticipate that abortion will provoke hostility and demands of punishment. That is, as we saw in section 1.3, it is one thing to believe that something is accepted, but another to believe that it is acceptable. So far, I am appealing to common intuitions to try to show that this distinction is real. However, there are solid empirical grounds for believing that such a distinction is involved when human beings make moral judgment. Psychologists Kevin Carlsmith, Paul Robinson, and John Darley (2002) tested which of two competing "philosophies of punishment" the folk generally clung to. One model they tested were the *deterrence* model, like the model adhered to by our imaginary creatures mentioned above, were we punish for the good consequences that follows. The option is the *retributive* (or the *justice*) model, the model currently under consideration, in which we punish because of a wrongdoing and the wrongdoer deserves to be punished. Initially, when presented with the two models, the subjects generally reported "a positive attitude toward both" and "did not display much of a tendency to favor one at the expense of the other" (Carlsmith et al. 2002, p. 294). However, when the subjects were given the task to allocate and distribute the punishment in terms of severity or guiltiness in response to a specific wrongdoing, the subjects were chiefly guided by "a just deserts motivation" (Carlsmith et al. 2002, p. 289). While we might have different reasons for expressing support for different justifications for punishment, our psychologies are quite naïve when things get specified. When presented with specific cases

[15] Cf. the first paragraph in section 1.4: here a *moral* creature is one that *morally judges*, not necessarily one that is *morally praiseworthy*.

people are consistently "driven by a strictly deservingness-based stance" (Carlsmith et al. 2002, p. 295).

The fact that moral judgments seem to imply desert might be the start of an explanation for their apparent inescapability. Our attitudes toward punishment and moral judgment both seem to be categorical in nature: just like the moral judgment is not contingent on our desires, deservingness isn't contingent on the consequences of the punishment. In arguing that when judging morally, we seem to be implying a notion of desert, I am not yet making any claim about the specific role natural selection has played in shaping our psychologies to be sufficiently naïve to allow for the categorical nature of our attitudes toward punishment or toward moral judgments more generally. That will be pursued in the next chapter. However, a more urgent matter is in need for explanation. Natural selection is an extremely conservative process; evolutionary biology consistently shows that we should expect it to "opt" for the simpler, cheaper solutions to adaptive problems even if they are less effective than more sophisticated alternatives (James 2010, p. 85). On that note, it might be thought that in so far as natural selection plays some role in our capacity for making moral judgments, we should be biased toward the deterrence model which seem to require less cognitive sophistication than the retributive or justice model. Because the former only requires the ability to recognize social regularities – in this case that some actions are followed by punishment. While the latter model, in addition to detecting regularities, requires us to recognize whether or not the punishment was justified. It is important, then, to account for the possibility of the sophistication of the mechanisms driving the moral judgment. I have argued that our capacity to *feel* is crucial to the evolution of human beings' cooperative traits and by extension to our capacity for moral judgment. Now I want to explicitly deal with the importance of the emotions and especially the moral emotion of guilt. I am going to claim that certain emotions are conceptually rich and cannot be entirely lacking in cognitive features.

Consider the relation between guilt and desert. Without desert there cannot be guilt – for what is feeling guilty if not judging that I *deserve* a retribution for my action? Moreover, without the moral emotions, like guilt, I doubt the possibility that we could ever fully grasp the notion of a *prohibition* – which we have seen is crucial for the moral judgment. To see this, it's expedient to compare ourselves with other animals. For all our similarities, it seems as if there is a crucial difference between human beings and other animals in terms of the cognitive abilities affiliated with the moral judgment. Again, take the difference between recognizing social regularities and

recognizing justification. Could our closest evolutionary relatives, for example, grasp this distinction? Primatologist Frans de Waal thinks they come close but not all the way. He offers a finer grained distinction of that between *descriptive* and *prescriptive* social rules (de Waal 1992). Descriptive rules correspond to our former distinction and denotes mere regularities in an organism's response to its conspecifics. This concept is not cognitively demanding and could be applied to every organism with observable behavioral adjustments to conspecifics; for example, fish and insects. Prescriptive rules, on the other hand, are also regularities, but they are not simply followed but *respected "because of active reinforcements of others"* (de Waal 1992, p. 244). De Waal asks us to think of the rule of avoiding a mother's aggression when defending her offspring. That rule, he says, goes from descriptive to prescriptive when *"members of the group learn to recognize the contingencies between their own and the mother's behavior and to act in a way that minimizes negative consequences"* (de Waal 1992, p. 244). Being ascribed the ability to grasp prescriptive rules it suffices to have a sense for *"a set of expectations about the way in which oneself (or others) should be treated and how resources should be divided, a deviation from which expectations to ones (or the others) disadvantage evokes a negative reaction, most commonly protest in subordinate individuals and punishment in dominant individuals"* (de Waal 1992, p. 242). De Waal is confident that chimps possess this sense, but does it follow that they therefore can be aware of "good", "bad", "just" and "unjust"? I think not. Just because they can recognize the difference between accepted and unaccepted behavior, they still fall short of grasping the whether the behavior is justified from a moral perspective. That is, they cannot think of the behavior as *acceptable* or *unacceptable* in general because, as de Waal himself points out, it is the dominant individuals that determines what's acceptable and not. But as we have seen, the moral judgment is categorical and therefore counts irrespective of the desires and inclinations of individuals, including whoever is in the dominant position within a group.

Although chimps have very sophisticated social cognition, this particular *authority-dependent* nature of moral judgments cannot be ascribed to them. When de Waal says that chimps have a sense for a set of expectations about how one *should* be treated; how resources *should* be distributed and so on, it's important not to conflate different interpretations of the word "should" or its relative the word "ought". It is an interesting question whether animals like chimps could be ascribed the hypothetical ought. Like we saw in section 1.5, the hypothetical ought is dependent on the subject of the prescription having the relevant desires or inclinations. Thus, for chimps to be able to make hypothetical imperatives, they must be able to form beliefs about

individual's desires. The mental life of animals is controversial territory and I am going to bypass any lengthy discussion of it. It's not important to my argument. However, at least they seem able to form beliefs about how desires can best be satisfied. Anyway, the point I want to make here, is that even though we are quite liberal in ascribing chimps with higher order normative mental states, it's a far cry to attributing them with a capacity to judge that something is deserved or that something is prohibited. I subscribe to de Waals own view on the matter: the chimps have some of the 'building blocks of morality', but not all of them. Their ability to recognize prescriptive rules is perhaps a necessary feature of morality, but not sufficient. In so far as they are our closest evolutionary relatives and have a lifestyle similar to our hominid ancestors, we should expect that what we observe in chimps is really the *precursors* of morality and that marks the difference between *proto-morality* and *morality.*

What should we say about the difference between the chimps and human beings that allows us to be attributed with the real thing and them not? One might be tempted to say that moral concepts are too *abstract* for other animals to grasp. But this is not at all obvious. Recognizing mother-offspring, differentiating between in-group and out-group, friend and foe and son on are all things that different animals seem to be able to do, which might permit us to ascribing them beliefs involving non-perceptual relational properties like sameness and difference. If it's not abstractness of belief that marks the difference, one might assume that moral concepts just are too *complex.* This too, however, seem to me not right. For as Joyce (2006, p. 81) aptly asks *"is the concept of prohibition really more complex than the concept of banana?"* If the distinct nature of morality involves it having some sort of rare complexity, it's unclear in what way. The relational concept of sameness and difference seem to me highly complex, and yet other animals seem to have it. Moreover, if it's not the abstract or complex nature of morality that explains why chimps and other animals can't make moral judgments, what does? I will argue that the explanation we seek is found in a consideration on the flexible cognitive sophistication of our moral emotions. This proposition is not uncontroversial; many people think that emotions lack the conceptual sophistication that moral judgments require and there are some observations that supports this conclusion. Since Darwin, it is widely thought that at least our basic emotions are adaptive mechanisms. They are each selected for by biological natural selection to do a certain job involving psychological, physiological in addition to behavioral elements triggered by environmental factors. They work independently of each other, each triggered by different but specific types of stimuli and ignorant of stimuli that is not of the right sort. Therefore, they are notoriously hard to "reason" with. What is usually called the autonomous nervous system

is a good example; in response to potential threats the autonomous nervous system has processed information and started relevant processes (fight, flight or freeze) long before you are fully conscious of what's going on. On this Darwinian picture of the emotions, it's easy to think that the emotions just lack the cognitive sophistication that we are looking for. I think that this thought is misplaced.

One of the emotions that is thought of as "basic" is disgust (cf. Ekman 1994). Darwin's idea was that at the core disgust has something to do with food rejection. Think about the fact that when we are disgusted, the distinct facial expression we get looks like we are trying to expulse food from the mouth or throat. In the extreme cases, the emotional response can evoke a gagging response even when what we are disgusted by is not being consumed. The psychologist Paul Rozin and his colleagues (2000), however, argue that disgust involved more than just distaste but also feeling contaminated and offended. For example, experiments have shown subjects resisting to wear a sweater that has previously been used by a stranger even when it had been washed several times. Subjects were even more hesitant when the previous owner of the sweater was believed to be a murderer. Responses to disgust is so strong that the emotion often generalizes to things that only *look* like disgusting things but are really mock-ups like plastic feces and chocolate that look like feces. Even though the subjects know that the objects are not what they seem, they feel disgusted by them. This leads Rozin et al. to the conclusion that animals (and infants for that matter) cannot feel bona fide disgusted because as the experiments shows, disgust involves the notions of invisible entities and the appearance/reality distinction which are considerable cognitive achievements lying outside the scope of the cognitive set-up of other animals. Although animals find certain things distasteful, by no means do we observe them exhibiting any extreme repulsion associated with disgust.

Before concluding this chapter, consider again the emotion of guilt.[16] So far, I have spoken much about other-directed moral judgments, but it seems fair to say that no moral system could ever work without a capacity for self-directed moral judgments. Just like we might assume that natural selection may have favored an unconditional altruism, we may think that an effective way of ensuring reciprocity and cooperation would be a mechanism that would kick in even if no one is watching (more on the advantages of guilt in the next chapter). Again, it's possible to

[16] I acknowledge that there are a several different philosophical approaches to the concept of *guilt*, for example the tradition of psychoanalysis etc. Thus, my claims and arguments about guilt should only be thought of within the context I am using them and does not necessarily hold for those other approaches.

think of guilt as just a certain feeling, but it's hard to see that there can be guilt without any kind of judgment that one *deserves* that feeling. Therefore, it is not surprising that psychologists associate guilt with moral transgressions (see for example Tangney 1992 for interesting discussion on moral transgressions as situational determinants of guilt). It is when one knows that one has done something wrong that the feeling of guilt appears. However, it doesn't seem to be a fear of punishment (in the formal sense of punishment) that sparks the guilt. Because you can fear punishment without being guilty, and you can also feel guilty without fearing punishment. If one is to feel guilty about breaking some rules, one must have had the capacity to somehow see those rules as justified overall and breaking them is unjustified. This is why the chimps observed by de Waal cannot feel guilty and by extension cannot make sense of a prohibition; their prescriptive social rules are only regulated by fear of punishment, while guilt requires the judgment that one deserves to be punished.

Chapter two: Is moral cognition caused by a genotype?

2.1 Is morality adaptive?

In chapter one I have hoped to have shown that our cooperative traits at one point in time (and perhaps still) is *adaptive*; that in the given environment they increase the fitness of the individuals. I have also indicated that its very likely that they are therefore caused by an biological *adaptation* that have been selected for by natural selection. Our capacity for altruism and other prosocial emotions are examples that could be part of our evolved bioprogram caused by a specific genotype. But these adaptations cannot be sufficient for our capacity for making *moral* judgments; morality is more than just prosocial emotions. We have not yet shown that the moral judgment is adaptive, and in turn, that would only be the start of an explanation of whether the moral judgment is an adaptation. These are the questions to which I now turn. They are quite pressing ones, and as we will see, have great metaethical implications. The question of whether the moral judgment is an adaptation will be treated as equivalent to the question of whether morality is innate or not. As I discussed in the introduction, an evolutionary interpretation on moral nativism leads to the evolutionary debunking of morality and not vindication. Naturally, then, a big part of the strategy for an evolutionary vindication of morality is to show that morality is not innate, which for our purposes means that moral judgment is not strictly an adaptation. Again, note that, although related to each other, this question is distinct from the question of whether morality is adaptive. It is necessary that adaptations are adaptive, but not sufficient; there must be additional evidence in place to affirm that certain adaptive traits are being caused by a genotype selected for by natural selection (see Gould & Lewontin 1979 for discussion on the scope of adaptationist explanations). But most explanations for adaptations must show how a trait is adaptive, so I will start there.

Once we have seen that cooperation is adaptive, we can easily imagine that morality too is adaptive due to its effects on cooperation. Some scholars even speculate that that the special features of morality that we identified in chapter one not only has a modest positive impact on human cooperation, but that it actually might *enable* the sort of relatively extreme level of cooperation we observe in modern human societies (See Kitcher 2011 ch. 2-3) for some discussion of this claim). Just think of the inescapable nature of moral judgements that I discussed in 1.5. That the moral judgments apply regardless of whether a person cares for it, at

least forces that person to recognize that something important that demands her attention has been expressed. Even if the person doesn't have the right inclinations, she cannot just ignore it, forcing her to engage with the demands of her peer irrespective of whether she agrees with the judgment from a moral perspective. In this way, morality forces us to permanently locate ourselves inside the cooperation game. It is extremely hard to imagine opting out of the game, for that is almost the definition of being a psychopath or immoral. Psychopathy is most commonly associated with the absence of the capacity for guilt or remorse. Remember from section 1.6 that guilt is an important emotion with a great deal of cognitive sophistication. When one knows that one has committed a moral wrongdoing, and one is not a psychopath, then we should assume that one is going to feel guilty. We say that such a person has a *conscience*, and it is upon the advantages a moral conscience that evolutionary ethicists often rests their case. I will return to the question of whether the evidence is strong enough such that we should explain the appearance of that phenotypic trait (the moral conscience) by reference to a specific genotype.

Why would a capacity to judge morally evolve by natural selection? The popular answer, which I hinted to above is that morality gives us motivations to behave in a fitness enhancing way. Since this is the cornerstone of every evolutionary account for morality, it's worth pausing over a long phrase from Joyce (2006, p. 109) himself:

> "*My own thinking on the matter is dominated by the natural assumption that an individual sincerely judging some available action in a morally positive light increases the probability that the individual will perform that action (likewise, mutatis mutandis, judging an action in a morally negative light). If reproductive fitness will be served by performance or omission of that certain action, then it will be served by any psychological mechanism that ensures or probabilifies this performance or omission (relative to mechanism that do so less effectively). Thus self-directed moral judgment may enhance reproductive fitness so long as it is attached to the appropriate actions"*

By "appropriate actions", we are of course talking about fitness enhancing actions, which in many circumstances will include cooperative behaviors. Therefore, to judge both one's own and other's prosocial behavior in moral terms, may be fitness enhancing. But if natural selection already has installed in us these prosocial emotions, powerful as they are, shouldn't that suffice? Certainly, on the surface, moral judgments seem fitness enhancing, but we could say the same

for magical superpowers such as being invisible. There needs to be a function that the trait is supposed to occupy for the theory to be credible. At this point it is often appealed to what is known as the phenomenon of weakness of the will. This is the observation that though human beings are very good at calculating what would benefit us in the long term, we are not so good at getting our motivations in line to obtain those benefits. Even If we know that short-termism is going to hurt us in the long run, we still capitulate to temptations. This is something most of us know from experience and there is plenty of empirical evidence that shows that weakness of the will is a marked feature of human psychology.[17] The idea is that when it comes to the realm of human cooperation, nature wouldn't want short-termism to prevent us from engaging in fitness enhancing behaviors, so it therefore opted for a special motivational mechanism for this realm – namely, the moral conscience. If an action entails the thought that it must be performed irrespective if one likes it or not (that the person cannot *escape* it), it doesn't guarantee that the action will be performed but it certainly will diminish other options that figure in the persons deliberations. In this sense, moral principles functions as "conversations stoppers" to use Daniel Dennet's (1995) catchphrase; they could be invoked in a decision-making process to help people from endless debate, negation and reconsiderations. The point is, then, that as a practical matter, judging oneself in moral terms cancels out different courses of action that figure in our deliberations in a manner that prudence doesn't.

So far, I think we have a perfectly good case *why* morality under some circumstances will be adaptive, but it's still not enough to show *how* it goes about being adaptive. We may still wonder what exactly has been done with the human brain to bring about moral thinking. Of course, we are in no position to lay out the neurological and (or genetic details if one is a moral nativist), but currently available evidence does lead us in a certain direction. It's without a doubt that the emotions play a crucial role in moral judgement, and in so far as evolution has been involved in forging our capacity to judge in moral terms, we should expect that its role has something to do with modifying our emotional architecture. Antonio Damasio's (1994) famous studies on what he calls 'acquired sociopathy' reveals how our emotions most likely play a necessary role in our moral judgments. Damasio studies people that for some reason have damaged their ventromedial prefrontal cortex – the part of the brain that processes emotions. His findings reveal that when people have their emotional capacities severely reduced, they show abnormal inability to act on reasoning – even prudential reasoning – and especially regarding social

[17] In behavioral economics this phenomenon is often called 'the present bias'.

decisions. For example, they lie for no reason and fail to express any regret or remorse for their misconduct. This suggest that a lack in emotional capacity parallels a lack in *moral* capacity. Research on psychopathy backs up this picture. Psychopathy is associated with an emotional deficit, but not prudential deficit (Blair 1995; Blair et al. 1997). In so far as psychopaths engage in abnormal moral behavior, it seems that the deficit in emotional capacity correlate with a deficit in the capacity for moral judgment. Moreover, even more convincing evidence has been brought forward from fMRI studies. When subjects were asked to morally judge on difficult cases, they showed marked and distinct activity in the brain's emotional centers (Greene and Haidt 2002). Therefore, we should not by surprised that research on neuroimaging on subjects playing out rigged game-theoretic scenarios (like the prisoner's dilemma or the ultimatum game), also report heightened activity in emotional centers when subjects choose to punish defectors (Sanfey et al. 2003).

Considering this evidence, one cannot overstate the importance of the emotions to moral judgments. In fact, it seems as if taking away the emotions from the equation means no more moral judgments. Especially the research on psychopathy show this. With their lack of a capacity for feeling guilt – a decisive moral emotion – they are simply unable to distinguish between *moral* transgressions and *conventional* transgressions (Blair 1995). Moral psychologist Jonathan Haidt (2001) convincingly shows, that not only is the emotions straightforwardly necessary for making moral judgments, they are in an important sense very often *causally* in the driving seat. What happens is that in identifying a moral transgression we typically have an emotional response *first*, which is followed by the reasoning concerning why the transgression was wrong were this reasoning looks like a kind of post hoc rationalization of the emotional response. As Haidt (2001, p. 828-829) himself admits, the causal chain might not always go this way, recognizing that sometimes we can start with reasoning that something is wrong and end up with a negative emotional response (The issue of moral reasoning and moral judgment will be further discussed in section 3.3). But such instances are rare and require special explanation, the default moral judgement is a post hoc rationalization of an emotional response. The hypothesis under consideration, that in the process of evolution our cognitively rich emotional systems have installed in us a capacity of a conscience which granted our ancestors reproductive advantages is still firmly on the table. Note that it's only a genealogical claim: its emergence is chiefly to be explained by reference to our emotional architecture and is entirely consistent with us having developed other means of arriving at moral judgments.

We now have a perfectly good explanation for why morality would be adaptive and what might have been done to our brain to make room for moral thinking. However, even though natural selection has undoubtedly played a crucial role in enabling us to be moral creatures, that doesn't automatically entail moral nativism. It may well be that the relevant modification of our emotional architecture is appropriately explained by the influence of culture on the moral psychology of human beings. Where the influence of culture is added to an existing psychology hardwired for cooperation which is more appropriately explained by reference to biological natural selection rather than *moral* thinking itself. This is the question I will shortly begin giving an answer to.

2.2 Taking stock: what is moral nativism?

Remember from the introduction that Joyce's evolutionary debunking argument depends critically on moral nativism. The reason evolution allegedly debunks moral knowledge is precisely because the capacity to recognize facts about moral reasons didn't evolve for that purpose. Thus, although we feel like our recognizing moral reasons are genuine, we are in the end making a fundamental error. But as we have seen in the previous section, that error is very useful and highly adaptive, so the debunkers case is plausible if moral nativism is true. That is the question I will now turn to.

First, we need to get clear on what we are looking for. What is it, exactly, that is supposed to be innate? As we have touched upon, the hypothesis we are going to be analyzing is whether *"morality (under some specification) can be given an adaptive explanation in genetic terms: whether the present-day existence of the trait is to be explained by reference to a genotype having granted ancestors reproductive advantage"* (Joyce 2007, p. 1). Joyce thinks that this amounts to granting human beings with a *moral sense* biologically selected for in the course of evolution, that provides us with innately specified information enabling us to *think* in moral terms. Specifically, it is a faculty for making moral judgments, where moral judgments are defined in a similar manner as in chapter 1; as distinct from prudential judgments. To get clear on how we should treat Joyce's nativist proposal, I suggest that we treat this hypothesis in the context of (Darwinian) evolutionary psychology which seem to be Joyce's preferred approach (Joyce 2006; 2016). This means thinking of moral cognition as supported by a specialized leaning mechanism or a "module" for morality. This need not involve all the assumptions of Fodorian modularity such as informational encapsulation etc. (Fodor 1983). Rather, I will

understand moral cognition in terms of *domain specific* innate psychological mechanisms, where these mechanisms should be explained by reference to a genotype (adaptation). This is, for example, exegetically supported by statements such as: *"human moral thinking is governed by dedicated mechanisms that evolve through the process of Darwinian selection"* (Joyce 2006, p. 17). Domain specificity is, of course, a somewhat controversial concept and is used differently by some scholars. I will delineate my usage of the notion with how it figures in nativists arguments, especially poverty of the stimulus arguments (These arguments will be discussed explicitly in the next section). According to Fiona Cowie (1999, p. 37), nativists appeal to domain specific mechanisms to provide an explanation for the gap between informational inputs provided by experience of the environment about some domain and the beliefs we obtain about that domain. Hence, I will treat domain-specific mechanisms as mechanisms that are not general empiricists learning mechanisms, but rather functions to perform special tasks.

In taking this approach, there are some preliminaries I want get out of the way before we can discuss the core issue. First, approaching nativism in this way means that one cannot simply undermine the nativist position by attacking genetic determinism. By reference to a domain specific cognitive structures, Joyce and other nativists who pursue this route, easily avoid the fallacies of genetic determinism because these mechanisms are entirely dependent on *specific* types of input. If this input is not provided by the environment, then the trait in question will simply not emerge. If our modern living conditions is sufficiently unlike that of our ancestors, there might be no society with any moral systems nor a single moral human being, but the nativist claim would still be theoretically possible. Secondly, nor is Joyce's approach threatened by the implausible idea that there are some innate moral principles. For it is not the content of moral judgments that is hypothesized as innately fixed but rather that one could make the moral judgment at all. Thus, Joyce's position is at least in principle compatible with there being an extreme amount of variation regarding what different cultures and people think of as morally right and wrong. As long as they actually think that *something at all* is morally right or wrong is enough for there being an innate moral faculty for making moral judgments. However, it's not plausible, I think, to hold that such extremely specialized mechanisms, hypothetically forged to solve some particular problem in the ancestral environment, could be entirely indiscriminate to the output they help generate. As Joyce (2007, p. 2) himself admit, if the moral judgment itself is an adaptation, it's probable that natural selection took some interest in favoring some broad and general principles too, given that they resulted in fitness advancing

actions. Interestingly, as Owen Flanagan's (2016) impressive anthropological overview shows, most of the moral norms we find in different cultures we find in others too, but they are thought to have different weight depending on the culture. But once we browse for possible candidates for universal moral principles, the environment is sufficiently rich and variant that an appeal to innateness is redundant. I will say some more on this before moving on to the core issue.

Do we have some candidates for moral universals? Its identification is perhaps itself a contentious topic. However, I think we have some natural candidates. When we see the suffering of others, most people tend to have strong affective response on behalf of the victim, especially if the inflicted suffering happened to an innocent person. Is there a universal prohibition against harming innocent people? I think that the evidence is weak. As the rampant anti-nativist Jesse Prinz (2007, p. 372) writes:

> "*torture, war, spousal abuse, corporal punishment, belligerent games, painful initiations, and fighting are all extremely widespread. Tolerated harm is as common as its prohibition. There is also massive cultural variation in whom can be harmed and when. Within our own geographic boundaries, subcultures disagree about whether capital punishment, spanking, and violent sports are permissible*".

Of course, most cultures have many rules against harming others, but these can be explained by socio-cultural factors and not biological ones. For example, harming others are often pointless in itself and we seldom gain anything from it. On the contrary, its often a bad strategy for maintaining a stable society. Thus, prohibitions on harming others might be explained by it being a precondition for social cohesion. One could respond by arguing that the innate capacity is the distress we experience when seeing harm being inflicted on others. Doesn't that in itself show that we are predisposed to oppose harm? I am actually sympathetic to this reply. Consider the case for mirror neurons mentioned in section 1.1, these neural mechanisms might be exactly the capacity in question here and it is most certainly in the domain of being granted Darwinian status. However, this does not mean that mirror neurons are *moral* neurons. Other animals, for example squirrel or deer, doesn't go around constantly killing each other. It doesn't follow that we attribute this behavior to a moral rule against killing. It's not necessary for them with such rules because they aren't biologically predisposed to kill conspecifics. Humans may equally not have such predispositions, but it doesn't follow that we are nasty creatures without such innate prohibitions. Perhaps unlike other animals we have a capacity for rational reflection, which we

could use to arrive at the conclusion that harming other may have positive outcomes. Therefore, we risk the presence of harm and violence in our societies. But it is this risk, not our bioprogram, that pushes through the construction of norms against harm. There are of course other contenders for innate moral rules. Joyce (2006, p. 65) mentions fairness, reciprocity, sustaining status hierarchies and regulating sex relations as plausible candidates, which universal presence he thinks indicates moral nativism. But all of these have dazzling variations across both space and time. In the modern world we have instances of cannibalism, slavery and headhunting; grotesque inequity; strongly egalitarian societies and rigid class and caste hierarchies; societies with extremely strict moral rules governing the body, while others are extremely permissive (Prinz 2007). Thus, it is not at all obvious that such rules are universal. Be that as it may, the bigger problem for this argument is that it only postpones the question of whether the construction of these rules is driven by culture.

Thus, relying on moral universals does not strengthen the case for nativism. On the contrary, I think it rather indicates that socio-cultural factors play the relevant role. Undoubtedly, from one perspective, there are great similarities between most of the worlds moral systems (cf. Flanagan 2016), and this does indicate that these systems have related origins. But that origin are at the very best equally well explained by reference to culture, and culture is certainly a universal. Culture, even though a universal, we know adds much more complexity and can easily explain any extreme variation in moral principles. For the innate moral universals view to be credible, far greater similarities should be prevalent. Consider the fact that in many parts of the world women are ascribed the same moral rights as men, while in some Arab countries killing a woman who has had sex outside marriage is not only morally permissible, but morally *obligatory* (Hauser 2006). It's not plausible that there is a genotype explaining the emergence of moral universals, and Joyce's position is strengthened by not raising this view. However, while worth mentioning, the moral universals view is not our real concern here. Because Joyce's main claim, which still would indicate debunking if successful is not that there are moral universals, but "*…whether having a system of moral judgements"* itself *"…is a human universal*" (Joyce 2016, p. 133). It is hard to find any group of people that don't have a moral system. Some have suggested that the Ik group of Uganda infamously characterized by anthropologist Colin Turnbull (1972) might be a candidate for such a group due to their alleged "viciousness" and sadistic customs. Turnbull describes a culture pressured into extreme individualism for survival and openly airs his horror at witnessing total disregard for family bonds, leading to deaths of children and the elderly by starvation. But even if a group seems

vicious, that fact is compatible with them having a system of morality. It might be a moral code that seems alien to us, however, a moral code nonetheless. Eventually, when the Ik elders heard about how Turnbull had described them to the rest of the world, they were outraged that Turnbull had destroyed their reputation and threatened to make him eat his own feces if he ever showed his face again. The fact that they were able to *judge* Turnbull to be worthy of such blame, shows that the Ik are capable of operating and employing moral concepts. Thus, Joyce's hypothesis of nativism about moral judgment is perfectly plausible. Yet I will argue that there is a better explanation for the origins of our capacity for moral judgments. I'll turn to that now.

2.3 Nativism about moral judgment

Without any evidence to the contrary, it seems to me correct to assume that all groups of people we know of possesses the trait in question, the one that we have been outlining in the former chapter – namely, a capacity for making moral judgements. Can we infer from this that such a capacity is an adaptation by natural selection, and ipso facto that moral judgments are innate? The apparent universal presence of the trait certainly lends nativism support, but additional arguments must be given in its favor. Note that we can distinguish two different but interconnected arguments working in the background – one inductive argument and one abductive argument. The inductive argument is this: all groups of people so far observed have a system of morality, therefore all future groups (given the right stimuli is provided) will have it as well. This is prima facie plausible. Morality goes further back than we can trace. As Joyce (2006, p. 135) writes "*moral precepts are mentioned in the Egyptian Book of the Dead and in the Mesopotamian epic of Gilgamesh*". Even more impressive is the fact that in so far as trade suggest an awareness of *ownership*, which in turn entail a sense of *rights*, we find the physical footprints of morality going at least back to the early Upper Paleolithic (see Mellars 1995, p. 398-400). According to Joyce, there is not a shred of evidence that morality is an artifact which emerged as a cultural phenomenon in 'modern civilization'. Like language, he makes clear, morality is ubiquitous and ancient. However, this isn't very convincing by itself. For what is also ubiquitous and ancient? Answer: culture. In due course I will show how it's perfectly possible to provide a competing 'how possibly' story about how something like morality might have grown out of socio-cultural developments. The inductive argument only has bite if the abductive argument succeeds. The abductive argument is this: the best explanation for the universal presence of morality is that there is a genotype causing the trait to emerge. Again, this

argument can be contrasted with the hypothesis that the true origins of morality is socio-cultural. Because we can provide an alternative story about how morality gradually grew out of a socio-cultural origin, coupled with the inevitable emergence of culture (perhaps itself caused by some genotype(s)), once morality was properly in place one cannot opt out of it. Therefore, morality would continue to reproduce itself into the future. One could complain that this would entail that some nativism must be true, perhaps a genotype(s) for culture. However, this would fail to establish *moral* nativism. Very few methodological naturalists would deny that our capacity for making moral judgments at least partially depends on some innate structures. As I will argue below, my own view is that we have specialized mechanisms in place that is *necessary* for making moral judgments, however, they are not *sufficient.* It's up to the moral nativist, then, to convincingly show that the gap between our innate predispositions and the environment is too big to be bridged by cultural factors such as socialization and learning. They must show that there is sufficient *poverty of stimulus*, such that by inference to best explanation, our capacity for moral judgement is innate.

Poverty of the stimulus (POS) arguments were first presented by Noam Chomsky in his attempt to argue that we have an innate language acquisition device, but it can also be deployed to argue for innateness in other domains such as the question currently under investigation here: are humans born with innate predisposition for a moral sense? The general idea behind the argument is – like many influential arguments in philosophy – quite simple. Suppose that the anti-nativist thesis is that we learn by applying domain-general learning mechanisms (as opposed to domain specific mechanisms), say hypothesis testing, to environmental input (stimulus). However, POS arguments holds that there just isn't enough information contained in the environment for the children to acquire the linguistic competence they exhibit through learning (Laurence & Margolis 2001). As a consequence, at least part of the explanation for the competence they exhibit, can't come from environmental stimulus. Furthermore, it follows that the earlier in development children acquire the capacity, the stronger the argument for innateness becomes (see Samuels 2002 for discussion of this claim). Joyce follows fellow moral nativist and philosopher Susan Dwyer (1999) in marshalling evidence from developmental psychology to argue that its simply implausible that the moral competence of young children is the result of an application of a domain general inference from environmental input. They draw attention to the robust findings of research on children's striking ability to recognize the distinction between moral violations and conventional violations. Consider canonical examples of moral violations (e.g. hitting, pulling hair) form conventional ones (e.g. wearing shoes on

the carpet, talking during storytime). The research shows that from a very young age, between 2 and 3 years, children are able to distinguish between such canonical examples on a variety of dimensions (See Smetana et al. 1993 for review of the psychological literature). Even more impressively, this also a cross-cultural phenomenon (see for example Nucci 2001 ch. 6). Psychologists largely follow Elliot Turiel's (1983) interpretation of the distinction were moral violations is thought to be more serious, more generalizable, authority-independent and justified differently. In an excellent review of this literature, Shaun Nichols (2005, p. 356-357) summarizes the findings nicely:

> *"…children tend to think that moral transgressions are generally less permissible and more serious than conventional transgressions. Children are also more likely to maintain that the moral violations are "generalizably" wrong, for example, that pulling hair is wrong in other countries too. And the explanations for why moral transgressions are wrong are given in terms of fairness and harm to victims. For example, children will say that pulling hair is wrong because it hurts the person. By contrast, the explanation for why conventional transgressions are wrong is given in terms of social acceptability — talking out of turn is wrong because it's rude or impolite, or because "you're not supposed to." Further, conventional rules, unlike moral rules, are viewed as dependent on authority. For instance, if at another school the teacher has no rule against talking during storytime, children will judge that it's not wrong to talk during storytime at that school; but even if the teacher at another school has no rule against hitting, children claim that it's still wrong to hit."*

This impressive level of moral competence from such an early age demands an explanation. How could these children possibly accomplish this? The findings present the moral non-nativist with a formidable challenge: if their hypothesis is to be credible, there must be sufficient learning and socialization between birth and the age of 2-3 years. To make things even more difficult for the opposition, Joyce and Dwyer points out that it seems that there isn't much explicit instruction for the young children to work with either. Normally, adults do not articulate the distinction between the moral and conventional verbally or any other way. The children get penalized for transgressions of both moral and conventional rules, which suggest prima facie that it would make it harder for children to recognize difference between them. Moreover, parenting styles, methods of socialization and other forms of explicit learning mechanisms vary greatly across cultures, but somehow children from all parts of the world seem to be able to

make the distinction. (Dwyer 1999, p. 171-177; Joyce 2006, p. 133-140). The moral nativists naturally draw two conclusions from these considerations – one positive and one negative. The negative conclusion is that the environmental information is sufficiently impoverished and unable to explain the child's moral competence. Joyce (2006, p. 137) writes:

> *"it is exceedingly unlikely that across the wide variety of human social ecologies there is some stable exogenous characteristic that may be plausibly appealed to as the explanans of this exceptionally regular ontogenetic sequence that characterizes the moral development of the human child".*

But the distinction is nonetheless made, thus the positive conclusion is that there is moral competence, and by inference to the best explanation, Joyce thinks that it must depend on innate domain-specific information about the moral domain. Ultimately, this must mean that the child has innate moral knowledge which is propositional in nature. The child can only make the distinction if she *knows* what morality in general is like. It doesn't follow that she knows what is ultimately good or bad moral rules or which principles that are in fact moral as opposed to immoral etc. That is, the child doesn't have any innately specified propositional knowledge about the *content* of particular moral systems. Neither does it follow that the child necessarily must know the moral domain from birth, not all design features are intended to appear then; just consider the case of puberty. All that follows is that knowledge of the moral domain is innately specified enabling the child to make distinctively *moral* judgments.

It must be admitted that the case for nativism is strong based on the evidence from developmental psychology on the moral/conventional distinction. But there are still room for considering other possibilities. For example, the opposition could still gather evidence that might indicate that children as young as 2-3 years receive the appropriate type and amount of instruction for them to master the distinction. As Prinz (2007, p. 387) argues, adults and parents use different reasoning patters to ground moral and conventional rules and children can learn to differentiate them by "…*imitating and internalizing the different reasoning patters in their moral educators*". There is empirical evidence that parents use different disciplinary tactics on different kinds of rules that the child violates (Smetana et al. 1993). Violations of moral rules tend to be enforced by power assertions and parents usually appeal to rights to apply them. Violations of conventional rules, however, are enforced by reasoning and appeals to social order. Nucci and Weber (1995) observe that children as young as 3 years are exposed to such

differences in rule enforcement and note that it most likely happens before that age too. This needn't be a strictly western phenomenon. Even if there are huge differences between cultures in parenting styles and methods of socialization, given that every culture has a moral system (i.e. that they operate with a distinction between moral and conventional rules), we should expect them to use differential rule enforcement in their children's early upbringing. This response to the poverty of stimulus argument presented by Joyce and Dwyer are rough and is far from a knock-down defense against the moral nativists. However, it does provide an alternative story adding plausibility to the notion that young children are able to *learn* the moral/conventional distinction. This story needs to be coupled with an explanation of the moral judgment that don't refer to innately specified information about the moral domain to be decisive.

I think we find such an explanation by once more invoking the role of our emotions in our capacity for moral judgment. As we have seen in section 2.1, the evidence that the affective system and the emotions that are based on it are in some sense necessary to moral judgments are overwhelming. Psychological altruism, guilt, disgust and all the rest of it runs on our affective system. That is not to say that they are lacking in cognitive sophistication, but that the affective system is their 'source of power'. Likewise, its seems that the moral judgment too fuels on our affective system. One of the consequences of Huntington's disease, for example, is the weakening of disgusting emotions. We also know that people suffering from the disease show high frequencies of paraphilias, which implies that they refrain from seeing divergent sexual behavior as wrong (Schmidt & Bonelli 2008).[18] There is also some evidence that alexithymia, a subclinical inability to identify and describe the emotions in the self, is highly associated with what the researchers call "Machiavellianism", acting prudential or instrumental rather than moral. The researchers write *"In particular, Machiavellianism was positively associated with externally orientated thinking and difficulty in identifying feelings."* (Wastell & Booth 2003, p. 731). On a related issue; one of the defining causes of psychopathic behavior is the more or less strong deficit in guilt and other negative emotions. Interestingly, psychopaths notoriously fail to recognize the wrongness of their actions (Hare 1993). Which consequences does this have for their performance on the moral/conventional distinction?

[18] Strong anti-depressants work to dampen our affective system, relieving the depressed person from the strong negative emotions. In light of this, some have actually argued that one "side-effect" of Prozac could be a "loss in moral sensibility" (see Kramer 1993, p. 278 for this interesting claim).

Research on psychopathic adults and children with psychopathic tendencies are interesting in this regard. Renowned cognitive scientist and psychologist James Blair has found that such people perform abnormally on the moral/conventional task. They are prone to explaining why moral transgressions are wrong in terms of the social and conventional. They simply fail to see the difference. Children with psychopathic tendencies are more likely to judge moral transgressions as contingent on authority than other children with behavioral problems. To them, the wrongness of hurting others, for example by hitting them, is only acknowledged if the teacher said so (Blair 1995). Morality, on the other hand, seems to somehow be independent of authority in the sense that it doesn't depend on the subjective desires of the person making a moral judgment – in this case the teacher. By recognizing the authority of the teacher, they are conforming to the conventional rule that the teacher is the leader in the classroom. Of course, normally, the teacher should be recognized as such to ensure a functional class, but the point is that this is not a *moral* justification for the teacher's authority. Imagine if it was the teacher that was the one hitting other children. Most children might recognize this as morally wrong, overriding their respect for the authority of the teacher as the leader in the classroom. The children with psychopathic tendencies, however, would presumably recognize this as morally right too "if the teacher said so". By their failure to distinguish between the moral and the conventional it seems that we must assume that they simply are unable to make a moral judgment at all.

We are considering that the role of our affective system for making moral judgments might do the explanatory work necessary to make an appeal the innateness redundant. I have shown how deficits in our affective systems corresponds to deficits in moral judgments. Blair and his colleagues also found that psychopaths have weak responses to distress cues in others. It is widely believed that our responses to distress cues are mechanisms that are evolved adaptations to restrain intraspecies aggression. Therefore, we should expect that, normally, people will display an affect-response to distress cues in others (perhaps some of the same mechanisms are causing our mirror neurons to fire). Anecdotal information indicates this as a plausible suggestion: if somebody attacks another and the person being attacked drops himself to the ground in submission, we readily expect the attacker to stop. Psychopaths, however, when witnessing others in distress, they show abnormally low physiological response compared to non-psychopathic criminals and autistic children (Blair et al. 1997). Blair and colleagues think the psychopath's mechanisms to respond to distress cues are damaged and that explains their lack of response. Blair call this the 'violence inhibition mechanism' (Blair 1995). Moreover,

they hypothesize that a damage to a similar mechanism explains their failure on the moral/conventional distinction. Normal people, on the other side, will experience the mechanism's production of a negative affect that generates the moral judgment. If this is right, then the capacity for moral judgment can be explained without an appeal to innately specified information. Of course, and as I have made clear several times, there is still a crucial innate contribution to the moral judgment, but the innate contribution is *affective* and not *propositional* as Joyce and Dwyer suggests. Why is the difference between these two potential innate contributions important? After all, on both accounts, the moral judgment has some necessary innate contribution. In fact, the difference is quite decisive. Contrary to the propositional account, on the affect account there is no *distinctively innate moral knowledge*. By consequence, then, if the affect account is correct, Joyce's evolutionary debunking argument loses its main premise: that the *moral* sense is caused by some genotype forged by natural selection. Given the centrality of this striking difference between these two nativist accounts, I should say some more about it: what does this difference consist in and what kind of contribution comes from innate mechanisms if not propositional knowledge about the moral domain? As we shall see, the answer to these questions is also the start of an alternative explanation for how we come to have knowledge of the moral domain.

We needn't fully endorse Blair's account of the moral judgment to subscribe to an affect-based account (see Nichols 2002 for a critical assessment of Blair's account of moral judgment). It simply is implausible that something similar to the violence inhibition mechanism alone can explain the capacity for drawing the moral/conventional distinction. It seems that it can only explain why we have a negative affective response towards a specific violation. But the capacity to draw the moral/conventional distinction requires that we can generalize that a specific violation belongs to a certain class of nonconventional violations. To be able to identify some violation as conventional and another as nonconventional, the affective response needs to have some difference in character. But how do we *know* when one or the other is the appropriate response? It seems that there need to be a normative 'theory' that specifies a group of violations providing our affective system with the information of when to bestow certain violations with a distinctive nonconventional status. Thus, even if the moral/conventional distinction doesn't derive from innate moral knowledge, the tendency to interpret the moral domain as distinctive might still in an important sense be unlearned. For as Dwyer (1999, p. 182) herself plausibly argues *"…it is hard to see how the deployment of emotional capacities could facilitate children's grasp of the distinction between rule-governed behavior and accidentally-regular*

behavior". Thus, even if the scope of the POS argument might not be sufficiently large to encompass moral knowledge, it might successfully lead to the conclusion that our capacity to acquire norms and to follow rules must be innate.

Consider again the distinction between hypothetical and nonhypothetical rules that we discussed in section 1.5. It's easy to see how an empiricist learner could use domain general capacities such as means-ends reasoning to learn from environmental input that certain actions get better results than others. She recognizes that she will get what she wants by following certain rules. But as we have seen, nonhypothetical rules count independently of one's desires. The very existence of rules of etiquette clearly shows that people have the capacity to acknowledge nonhypothetical rules even when they conflict with one's desires or inclinations. There is no obvious story to be told about how an empiricist learner might come to learn nonhypothetical rules. Moreover, the earlier we observe this capacity in humans, the hypothesis that we are designed to detect rules becomes even more plausible. We should be impressed, then, when there is solid evidence that children as young as four are able to do this (Cummins 1996). What more can we say about this capacity? The leading account is due to philosophers Stephen Stich and Chandra Sripada and pending more information it is the theory I myself subscribe to. They propose that we are innately equipped with something they call the 'norm-acquisition system' (see Sripada and Stich 2007; Sripada 2005 & 2008). For our purposes, we can think of norms here to denote some type of social rules prescribing what can and cannot be done. Contrary to Blair's more crude account, Stich and Sripadas theory is a dual system: there isn't just a single mechanism doing the job. Here is the system spelled out in pure functional terms:

> *The job of the Acquisition Mechanism is to identify the norms in the surrounding culture whose violation is typically met with punishment, to infer the content of those norms, and to pass that information to the Execution Mechanism, where it is stored in the Norm Data Base. The Execution Mechanism has the job of inferring that some actual or contemplated behavior violates (or is required by) a norm, and generating intrinsic (i.e. non-instrumental) motivation to comply and to punish those who do not comply. (Stich 2008, p. 228).*[19]

[19] This quote is not found in the work where Sripada and Stich outlines their hypothesis, but rather from Stephen Stich's review of Joyce's *The Evolution of Morality*. Oddly enough, that were I found the clearest statement of the mechanism.

Such a system almost certainly was adaptive in our ancestral environment. As Sober and Wilson (1998) showed in their theory of group selection, even though it is costly to enforce norms, it is even more costly to defect. The fact that our closest evolutionary cousins, the chimps as observed by Frans de Waal, display similar behavior renders the idea that we have an innate rule-detecting system with even more plausibility. It indicates that precursors of the system were already in place before we split into the more immediate human ancestry.

To conclude this crucial theme, we can happily accept the negative result of Joyce and Dwyer's POS argument: the environment is by itself sufficiently impoverished to explain young children's moral competence. What we have good reason to believe is that we have unlearned rule-detection system in addition to the innate affective system, like Sripada and Stich's norm-acquisition system, which helps us in this regard. But, as we have seen, we needn't accept their positive conclusion that we must be innately endowed with domain specific knowledge about the moral domain. For the capacity to comprehend rules are by no means a distinctively *moral* capacity. There are some final reasons why I think this nativist option is more plausible than *moral* nativism. First, one advantage of rule nativism over moral nativism is simplicity: it is much simpler and less demanding than moral nativism. We know from before that natural selection is an extremely conservative process, which is something that Joyce (2006, p. 22, 114, 115) himself repeatedly concedes. Time and again, we learn from evolutionary biology that natural selection will "prefer" simpler and cheaper solutions to adaptive problems even if they are less effective than more sophisticated ones. Secondly, a mind equipped with norm-acquisition device will generate moral judgements that are uniquely adapted to its cultural environment and will more readily explain the worlds moral diversity as discussed in the previous section. Although moral nativism is compatible with moral diversity, rule nativism sits more comfortably with it. Third and lastly, even though there are extreme diversity in moral practices, there is a more general sense in which there are similarities: we find some of the same moral values in every culture we have so far observed. Anthropologist Oliver Scott Curry and colleagues in an impressive ethnographical study of 60 different societies testing the hypothesis "is it good to cooperate?", found that there are seven moral rules shared by every society. They are 1) help your family, 2) help your group, 3) return favors, 4) be brave, 5) defer to superiors, 6) divide resources fairly, and 7) respect others property (Scott Curry et al. 2019). This also suggest, I think, that there is *some* stable characteristic that together with our innate capacities account for our moral competence. But if this isn't to be explained by reference to another

fellow genotype, what else could there be? There is one more thing that all those 60 societies had, and probably every society that ever was had in addition to genes: a set of social conditions that we might call *a culture*.

2.4 The role of culture in shaping *social* cognition

In showing that moral nativism doesn't hold up against scrutiny, Joyce's evolutionary debunking argument has taken a serious hit. Remember that the debunking argument depends on biological evolution to do all the explanatory work involved in accounting for our capacity to make moral judgments. If moral nativism is true, then we should understand our capacity to make moral judgements by reference to a genotype causing the trait to emerge. If this hypothesis is supposed to be credible, that genotype must have been an adaptation selected for by natural selection. But as I have hoped to have shown in the last section, even if moral nativism could in principle explain our capacity for moral judgement, there are theoretically better options on the table. This is the first step in arguing that evolutionary theory might be deployed in attempts to vindicate our moral commitments, rather than debunk them as Joyce thinks it does. However, as it stands we have only a negative conclusion. Refuting moral nativism does nothing more than *neutralizing* the dispute between the debunkers and the vindicators. We are still left explaining the distinctively *moral* contribution to the moral judgment. And we still risk undermining our moral beliefs if we can offer no alternative. It's important to note that moral nativism is logically distinct from other skeptical worries about the epistemic justification of our moral judgments. One can still argue that the nature of moral judgments is not compatible with evolutionary theory. That is, one can still hold an error theory in which our moral judgments necessarily cannot be able to do what it 'purports' to do: transcend institutions and apply independently of individual's motivations. But since we cannot fit such normative properties within the natural world, it seems that we make a fundamental error when making moral judgments. If moral nativism were true, then we have a solid explanation for why we make this error, but the argument could be made independently of it. Nativism is a genealogical argument, while the latter is a conceptual argument about moral judgments. Of course, nativism is extremely important positive argument for the evolutionary debunking of morality, while the conceptual claim is only a negative argument against the evolutionary vindication of morality. I will confront the conceptual claim head on in the next chapter. For now, it's important to note that we still need an explanation for manifest morality. It does not seem as if our innate capacities can explain the distinctively *moral*. The alternative thesis that something as diffuse

and vague as the notion of *culture* can do this explanatory work we need, will be born out from here onwards. Thus, this section marks a turning point in the argumentative thread of the thesis.

In the last section we saw that an appeal to our innate affective system renders the appeal to innate moral knowledge redundant. Furthermore, as Nichols (2005, p. 367-369) keenly observes, there are other interesting consequences which follows from how innate affective mechanisms shape our moral capacities: those mechanisms are neither domain specific nor domain general. To see this, lets first revisit Blair's (1995; Blair et al. 1997) research which suggest that the affective responses to suffering emerge early in humans and appear to be a cultural universal. Hypothetically, the affective responses produced motivations and inclinations that were fitness enhancing. This is only theoretical speculation but given that these innate affective systems are intimately linked with behavioral response, its plausible that they were designed by natural evolution. Thus, for the sake of argument, let's assume that they are the result of adaptations to a background problem in the ancestral environment. Moreover, the evidence from research on the moral/conventional distinction suggests that morality has the marks of domain specificity – which among other things, leads Joyce to entertain the thought that this distinction must come from within. Apparently, violations get organized into the distinct cognitive domains of the moral and the conventional. But I have argued that these domains are due to affective systems. Perhaps the ones that respond to other's suffering. But this example was exploited due to its diminished presence in people who struggle to grasp the moral domain – namely, psychopaths. The affective system, of course, are designed to do other jobs as well. Therefore, the affective system will impose structure onto *different* cognitive domains and would constrain them in ways that are not domain specific. Even if the affective response to suffering does help us to demarcate between the moral and the conventional, it's unlikely that only these specific cognitive states get affected by it. The affective clout influencing the nature of the moral judgment is most likely not specific to the domain of moral judgment but is crucial in other domains of knowledge as well. For example, *"…our response to suffering in others might also play an important role in the way we think about natural disasters that cause immense human suffering"* (Nichols 2005, p. 368). Even if these affective systems are not domain specific, this doesn't mean that they are domain general. There are lots of domains of knowledge that they don't influence such as our cognitive states about, say, physics, farming or reading. Rather, it seems as if these affective systems are domain *diverse*. This alters the picture about nativism we discussed in the former section. Because the nativist arguments that the environment is sufficiently impoverished might succeed without it following

from this concession that there are innate mechanisms devoted to the domain in question. But even if the moral judgment depends on domain diverse mechanism, how are they sorted into the specific moral domain? A mechanism for rule comprehension would get us going with norms, but not distinctively moral norms. I propose the thought that we use culture to classify the moral realm. In the following, I will end this chapter by presenting the cultural evolution of what has been dubbed the *ultrasocial* (as compared to "only" *prosocial)* nature of human beings. This will be the platform I use to develop an alternative evolutionary genealogy of morality in chapter 3. This alternative evolutionary genealogy, I argue, shows prospects of vindication rather than debunking morality.

The fact that morality depends on domain diverse mechanisms might leave us wondering if the range of structures that such mechanisms might impose on cognition ever comes in conflict with each other. That there is a risk of the system wrongly interpreting the cues in the environment generating inappropriate responses. There is probably a chronic version of this going on in psychopathy. However, the fact that we normally do have the appropriate response should leave us impressed with the precision of interplay between the information we find in the environment and our interpretation of it. Leaving aside the moral domain for now, we should be even more impressed to realize that even when specific domain diverse innate systems might conflict internally, different innate systems might themselves in some sense be opposed to each other. Alongside our inclinations to care for kin, our willingness to engage in reciprocal cooperative relations, and the possibility to act altruistically toward non-kin, we also have strong instincts for self-preservation. All these capacities developed because they enhanced fitness. Humans beings, then, have evolved strong motivational capacities to both be self-interested *and* have genuine concern for others. Given this profound prima facie ambivalence of human nature, how is it that we so successfully are able to stabilize the internal pressures and appropriately adapt and guide ourselves through the constant threat of destructive conflicting motivations? It follows that there is acute need for an external structuring. Following the anthropologists Peter Richerson and Robert Boyd, I propose that cultural norms play a large part in structuring our motivations. Even if the processes we looked at in section 1.2; kin selection, reciprocity, and group selection have resulted in the emergence of cooperative traits, the models of Richerson and Boyd (1985) suggest that they can only work for relatively small groups with large degree of shared genetic makeup. They would have to be something along the lines of a chimpanzee troop. This is a notable finding and demands further explanation of how cooperation have extended from groups of no more than 100 conspecifics to cities of over

30 million. Another interesting consequence of this is that in the case of morality being a genetic adaptation doesn't need such an explanation. As Joyce (2009) correctly argues, the processes of kin selection and reciprocal altruism could in principle be enough to explain morality as genetic adaptation. On this picture, morality could have evolved when our ancestors lived in relatively small groups. However, since we have seen that this isn't likely, we must look elsewhere for an alternative genealogy.

Let's pick up where we ended section 1.2. How can we account for the ultrasocial cognitive abilities of human beings enabling the kind of massive scale of cooperation we observe today? For example, we could imagine that one way for reciprocity to generate larger-scale cooperation would be the possibility of punishing non-reciprocators. If reciprocators could punish defectors, defecting might be tipped over to be fitness-decreasing behavior. In that way, the core group of reciprocators would not be overrun by defectors in the course of evolution. However, this should leave us wondering why those administering the punishments won't lose out in the evolutionary struggle to more easy-going reciprocators – those who reciprocate but aren't prepared to use energy on punishing others. However, the theory of group selection provides us with an intriguing answer: a *group* of punishers may readily outperform a group of easy-going reciprocators. Although being a punisher generally will come with a cost, the benefit to the group (and ipso facto to the individual), can easily outweigh the cost. Therefore, if a fitness-sacrificing trait have evolved by group selection, it must have been as the triumphant in the competition between the forces of individual fitness advancement tipping the scale one direction, and the counterbalancing of group-benefiting fitness sacrifice. One great achievement of Sober and Wilson's theory is to show how the balance need not always tip in the favor of individual fitness advancement. However, it's hard to imagine that group selection easily occurs at the genetic level. The reason is that it is very likely that intergroup mating probably was quite common in ancient times. Even if two or more tribes are in constant conflict with each other, the victorious tribe in potential warfare could end up taking the women of the conquered tribe. Notice that this only needs to happen once for it to start countering genetic selection. But this isn't a problem for the hypothesis here, for kin selection, reciprocity, and even genetic group selection isn't strictly theories about the psychological motivations of individuals, but rather outlines of evolutionary process that might resulted in the emergence of cooperative traits in relatively small groups with high degrees if genetic relatedness. The genotypes responsible for such traits are of course convenient building blocks for expanding cooperative traits, but they can't by themselves explain that expansion. But group selection needn't occur at the genetic

level. One of Sober and Wilsons celebrated ideas is that the notion that the units of natural selection might be bundled up together much like Russian matryoshka dolls. Genes compete within an animal, animals compete with other animals within a group, and groups compete with other groups and so forth. Remember that Darwin himself was ignorant about genes, nonetheless he articulated the theory of evolution by natural selection. As one will find in every textbook on natural selection, so long as there is trait variation, heritability and differential reproduction, there is selection. As far as I can tell, it doesn't necessarily follow from the theory of natural selection that the traits in question *must* be written in the genes, or that only individual organisms are the reproducing entities. A culture, for example, may produce the kind of traits we are after. If so, *cultural* group selection (cf. Boyd and Richerson 1985) may provide us with an explanation for why large-scale cooperation emerged.

The possibility of group selection depends on a sufficient degree of intragroup uniformity and intergroup variability. The widespread phenomena of intermarriage and migration are efficient obstacles for these criteria to be satisfied at the genetic level – however, they are more plausibly satisfied at the cultural level. I follow Richerson and Boyd (2005, see ch. 5) in thinking about culture as information capable of affecting individual's behavior that they acquire from other members of their species through teaching, imitation, and other forms of social transmission. Were information is taken to mean mental states that is acquired or modified by social learning and affects behavior. Consider the tendency of human beings to conform our behavior to that of the majority of one's group. There is solid evidence that conformity is a persistent trait of the human species. Boyd and Richerson (1985) argue, plausibly, that being conformist will tend to be adaptive in a variable environment because it allows reliable and efficient access to successful behavior. By emulating the majority, members of a group could be conferred an evolutionary advantage by adopting successful solutions to common problems in the environment previously worked out by other members. In this scenario, individuals needn't reinvent the wheel, but can instead follow "ways of doing things" eventually established as cultural norms over some period of time. In addition to mechanisms that support conformist transmission, there are a range of cognitive traits that plausibly lies at the heart of human's unique cumulative culture. Michael Tomasello (1999, see especially ch. 3) stress the importance of social learning through imitation of other's intentional acts. Imitative learning, he argues, also is a prerequisite for language acquisition – a prime driver for culture. Let us, for the sake of argument, think, with Tomasello, that we are innately disposed with dispositions and capacities to create culture. Together with traits of employing punishment strategies, our

capacities to learn and conform to the successful behavior of others, there is a natural explanation for why intragroup variation might be suppressed while intergroup variation increased. Given that the ones who punish are more successful than easy-going reciprocators,[20] new members reproduce such behavior by acquiring the norm of behaving in this way. However, there is a problem with the power of punishment. For punishment could in principle settle any sort of behavior within a group, even obscure behavior and behavior that is maladaptive. It could mean the sorry end of any group if the behavior it ensures within a group is bad enough for them to cope with their environment. But it could also result in a thriving group if the behavior is adaptive. This is the crust of cultural group selection. Once there are many culturally distinct groups, there is selective pressure for those cultural systems that are promoting "prosocial" or cooperative traits. In the long run, then, the groups with such cultural systems will outcompete groups with less successful behavior. To put it crudely, a group with a cultural system obsessed with wearing plants on their head would eventually be outcompeted by groups who stressed self-sacrifice and the welfare of one's fellow group members.

Here we have a perfectly plausible explanation for why large-scale cooperation might have evolved. Over hundreds of thousands of years with the right selective pressures ended up favoring individuals who were members of the groups with cultural systems promoting the traits enabling cooperation at that level. This isn't only plausible "on paper", Boyd and Richerson think that culture was originally an adaptation to the highly variable Pleistocene environment with its climatic chaos. The story is backed up by the fact that the long period of the Pleistocene overlapped considerably with the period in which humans started living in social groups with cultural institutions. In such environments, learning to cope is difficult and costly. As a storage of information about how to successfully cope, the emergence of cultural systems let individuals selectively use environmental cues to learn previously worked out solutions by imitation and teaching. Culture is adaptive because it can do things that genes cannot do alone. It can produce complex adaptations much faster than genes can do on their own. As Boyd and Richerson (2005, p. 146) write:

> "*When the kinds of social learning biases (imitating the successful) are combined with occasional adaptive innovations and content biases, the result is the cumulative cultural*

[20] Which needn't always be the case, but often are (see Boyd and Richerson 1985).

> *evolution of complex, socially learned adaptations, adaptations that are far beyond the creative ability of any individual*".

Therefore

> *"…cumulative cultural evolution gives rise to complex adaptations much more rapidly than natural selection give rise to genetic adaptations…*"

Moreover, when they "created" cultures, our ancestors massively influenced the ecological niche in which they lived, and as this is occurring, genetic individual selection might be ongoing, rendering the course of the latter process partially dependent on outcome of the former. The culturally evolved environments where prosocial norms are enforced by effective systems of punishment and reward, individual selection will favor psychological mechanisms suited to this environmental niche: individuals more likely to gain social rewards and avoid punishment. Thus, we might say that humans have evolved two different set of innate cooperative dispositions. One set of ancient capacities we share with our primate relatives, shaped by familiar processes of kin selection and reciprocity, and one set of group capacities enabling the large-scale cooperation we observe today. By constructing our own niche, our genes and culture coevolve, as is demonstrated by the extraordinary increase the human brain the last three million years. Without cultural innovations such as the use of tools, controlling fire and cooking etc. our hominid ancestors would never developed such brain size because the selective pressures in the appropriate environmental niche wasn't there.

Chapter three: Why morality is not an illusion

3.1 An alternative genealogy of morality

At the end of chapter two, I argued that human ultra-sociality developed through gene-culture coevolution. Cultural group selection explains how human beings eventually came to have more indiscriminate forms of altruistic capacities enabling cooperation on the scale we observe among human beings today. But this isn't yet a story about our apparent *moral* capacities. We are still working under the assumption that there is something distinct about morality which we can indeed observe in the behavior of human beings. Though manifest morality can be observed, it seems to involve, at least from a naturalistic perspective, some very strange properties. For example, moral judgments seem to be treated as inescapable in a way that conventional judgements are not. Again, we must ask: how is this possible?

In the former chapter we discussed a possible explanation for these strange features of morality: that there is a genotype forged by genetic selection shaping and supplying moral cognition with this specific information. When our hominid ancestors lived in relatively small groups, with inclinations to sacrifice for kin and perhaps others if one expected that the sacrifice would be reciprocated, weakness of the will would in many cases deteriorate cooperation. The solution is then to create a *conscience* operating in a novel *moral* domain, where properties such as "right" and "wrong" is recognized to exist independently of subjective desires and inclinations. By recognizing such properties, one cannot simply say "No, I don't want to do that" because they simply don't depend on what your or anyone else's existing motivations are. This 'oomph' involved in the moral judgment works by nudging us away from problems stemming from weakness of the will. This, in turn, explains the practical clout that seem to be imbued in moral judgments. Of course, all this is just a useful illusion imposed by the relevant genotype. Natural selection doesn't at all "worry" itself with moral properties, all it "cares" about is increased fitness, and creating the *appearance* of a moral realm with such properties does just that; it is fitness enhancing. As we have seen, this story depends on moral nativism, and although there is some evidence that support moral nativism, we have good reason to look for an alternative story. What if morality is rather a consequence of cultural evolution enabled by our ultra-sociality? That its distinctive features are not an illusion produced from within but is rather properties we can recognize in the exogenous

accumulative cultural environment we have created over hundreds of thousands of years. On this view the origins of morality are not due to a genetic adaptation, but a cultural adaptation and therefore its emergence has a quite different history. On the former genealogy, there is a reference to a genotype selected for in the course of biological evolution. On the latter genealogy, however, to which I now turn to, we need to tell a story without such a reference.

How did we get from there to here? How did we go from being psychological altruists to be psychological moralists? There is no doubt that such a process occurred. There is a point in hominid history where we weren't psychological moralists and there is a point – now, for example – were we are. Clearly, our knowledge of the specific details of this story is extremely incomplete. But there is not absolute darkness. We have enough data to say *something* about how morality might have developed from its precursors. In accordance with the general methodological approach laid out in the introduction we are only trying to infer from the evidence to the *best explanation*. Even if we don't know what *actually* happened, if we could construct a hypothetical outline of how it *possibly* did happen, at least it has gained plausibility. A leading hypothesis is represented by Philip Kitcher. He proposes a three-stage process starting off with conditions similar to that of our closest evolutionary relatives. We know that gibbons live in small family units, gorillas in small groups and orangutans relatively solitary. Only chimps and bonobos live in bisexual groups with both adults and juveniles. Studies of hominid remains suggests that the social life of our hominid ancestors was quite similar to contemporary life of chimps and bonobos. How can we explain the difference in sociality between chimp-bonobo-hominid and the rest of our evolutionary relatives? Kitcher generalizes research by primatologist Richard Wrangham into what he calls *the coalition game*. Wrangham proposed that the social structure of chimp-bonobo groups is determined by female foraging strategies. If this is true, Kitcher supposes that if we generalize the emphasis on foraging, we may recognize that competition among weak/vulnerable individuals would require their participation in coalitions and alliances. If animals go through a stage where they are relatively weak, a propensity to forming coalitions is likely to prevail. Responding blindly to the preferences of another animal probably does better than a complex calculation of future benefits, creating a selective pressure for a capacity for psychological altruism. However, psychological altruism is a fragile notion by itself. Just consider once again the ambivalent nature of human, or other primates, motivations. To gain insight into how this period might have looked like, we can observe contemporary chimp-bonobo tribes.

Frans de Waal (1989) argue that chimps and bonobos have a capacity for psychological altruism. In his research he observed motivations to reduce conflict within groups of apes and monkeys, such as breaking up fights but not choosing sides, which cannot adequately be explained by psychological egoism. He observed behavior such as giving care and relief to distressed to nonrelated individuals. For such behavior to be possible, there needs to be in place genuine concern for others and sometimes even the capability of understanding their needs and emotions. However, de Waal points out, even though this capacity is real, it rarely prevails amidst a highly complex and ambivalent motivational context. Psychological altruism is limited in all kinds of ways. The weight altruists give to their beneficiaries wishes can range from minimal gesture to total subordination. They may only respond to some individuals or group of individuals and not to others. When an altruist does respond, their alignment of preferences with that of the other can vary greatly from context to context. Human beings must have something else – something extra to work with. If not, our social lives would be very different from what they are. With a fragile capacity for psychological altruism, chimps and bonobos are able to live together in small groups, which is advantageous. But as Kitcher (2012, p. 307) points out, it doesn't permit them to co-habit with ease. de Waal's study of chimp and bonobo troops describe a social life that is regularly on the verge of breakdown. Tensions and hostile agitation emerge almost every day and intricate and time-consuming mechanisms for keeping peace are performed. In fact, they might use three to six hours every day mending the social fabric of the group. We must assume that once upon a time our ancestors faced similar conditions.

How did we overcome the predicament of this first phase? Kitcher (2011, p. 76) introduces the concept of an *altruism failure* to account for a context where the altruist does not align with the preferences attributed to the beneficiary. For this next step to be possible, human beings must have evolved an ability to override the temptation to do something that would constitute an altruism failure. We would have to have the capacity to formulate commands to ourselves and indeed act on that command – this means being susceptible to rule-following and rule-comprehension. Therefore, the persistent altruism failures create a strong selective pressure rule comprehension, eventually resulting in a genetic adaptation along the lines of one of the mechanism granted innate status in section 2.3; perhaps something like the norm acquisition system. In the earliest stages the capacity to follow rules replaces altruism failures with at least behavioral altruism. One thing that we know about contemporary human beings and our ancestors is that we have a remarkable capacity for pattern recognition. This ability

likely served rule comprehension well by perceiving regularities in actions and consequences. Our ancestors likely observed that following particular desires led to trouble. Avoiding these saved them from a lot of distress. But it's not only based on fear:

> *"It can be articulated, superseded and supplemented by a variety of other dispositions: a sense of respect or awe, a desire for social approval, solidarity with the group, and so forth. Human beings have developed a whole arsenal of techniques that encourage self-command. As with biological functions generally, it is well to have multiple back-up systems"* (Kitcher 2012, p. 309).

An ability for comprehending and following rules, would thus mean an enhancement of psychological altruism resulting in improvements in being coalition partners, bolstering the advantages of forming coalitions and alliances and expanding them into greater numbers.

With our new normative (but not yet moral) capacities, the group grew larger and the social arrangements more complex; we entered the next phase. This is where the development of morality, I suggest, became intertwined with cultural evolution and cultural group selection. The development of culture meant that we could socially embed our newly developed capacities to formulate and follow rules. We could now share commands and normatively guide each other. Studies of contemporary hunter-gatherer societies can give us a picture how the new normative lives of our ancestors were socially embedded. Anthropologist Christopher Boehm compared some of the few remaining contemporary hunter-gatherer groups from South-African !Kung-speaking foragers, to Inuit Eskimos and Navajo Indians. He observed that members, mostly adults, usually assemble at the "cool hour" to deliberate and they flesh out the scheme that should govern their social lives (Boehm 1999). These proceedings would have been very useful back in the harsh Pleistocene days. In such conditions, we can imagine that the contribution of each adult is crucial to the well-being and probably also the survival of the whole group. No member's wishes can be neglected. This creates the need to arrive at social arrangements acceptable to all. Kitcher supposes that those then new normative capacities have been socially embedded in this way for tens of thousands of years. For example, it may have emerged with the development of human linguistic abilities – that is, at least fifty thousand years ago but probably more (Kitcher 2012, p. 309)

However, the selective pressures to enforce cooperation never disappeared. Today, we still witness widespread failure to cooperate and align ourselves with the preferences of others. Weakness of the will type problems are consistently present and is most likely the price we pay for some other valuable end: our ability to calculate subjective preferences in a flexible way. Joyce (2006, p. 110) describes this internal conflict in this way:

> *"A creature that in any circumstances conceives of a banana as the highest good will be at a constant disadvantage to a creature that is able to reassess the value of the banana depending on circumstances. A banana isn't worth much when you are full; when you are on the brink of starvation it may in fact be the highest available good. But the inevitable price of any such plastic goal assessment (i.e., practical intelligence) is error. Just as biological natural selection cannot create a creature that has a flexible belief-formation system and always forms true beliefs, nor can it build a creature that has a flexible capacity for assessing the subjective value of things and always does so correctly"*

These types of practical error would have been destructive for cooperation, whose benefits are often long term. Actions such as reciprocation and protecting kin are both directed toward a future return on investment, just as refraining to eat junk food even when you have innate mechanisms telling you how good it really is. The solution to counter weakness of the will in the crucial domain of cooperation was morality – both Joyce and I agree on that. But contrary to Joyce, I think that the trait in question emerged out of cultural group selection. After our innate normative capacities became socially embedded, human behavior became profoundly shaped by the different paths of cultural evolution. This history can be seen as an extremely rich and diverse "experiments in living" where a vast range of socio-cultural ecologies lead to different outcomes. Groups formed different ecological niches where fitness enhancing behavior became "trapped" inside a positive feedback loop amplifying the development of the trait. The most successful experiments were taken up by other groups, as individuals and groups of individuals migrated and joined up with each other. Eventually, cultural group selection selected the group(s) with the best solution to the selective pressures and their decisive solution was the moral judgment. We are all children of those successful groups, and that's why we observe the moral judgment in every corner of the world. The question of whether there is a marked point in this story in which morality starts I think is misplaced. Because if this alternative genealogy is correct, there simply isn't *one* starting line. In one

sense, it all started with the genetic adaptation of prosocial emotions, in another sense it was the capacity to formulate and follow rules. Or it could be the first group who punished *some* kinds of violations with a bit harder than other ones, perhaps starting to make a distinction between moral and conventional rules. It hard to settle between them, because they are all correct. Rather, I think, with Kitcher, that what we are describing is a story of a human *project*, constantly developing in new directions and never finished. But because that project is an integral feature of human cultural evolution, we cannot opt out of it because we cannot opt out of culture.

I want to finish this section with a qualification. This view amounts to a version of moral constructivism. However, unlike rationalistic or contractarian accounts of moral constructivism the moral judgement is not an artificial innovation. It is as natural as any other product of human evolution. Just because morality isn't produced by a genotype, its equally natural. What I have in mind is perfectly captured by John Dewey's famous remark: *"moral conceptions and processes grow naturally out of the very conditions of life"* (Dewey & Tufts 1932, p. 343). Not all conditions of life are simply innate and ipso facto not all of human nature is essentially genetic.[21] This results in a novel statement of moral constructivism, because it leaves out any deliberate or intentional design. Our ancestors, the ones who "created" the moral judgment, of course, never really did so intentionally, or with any knowledge of what they were doing. They were just shaped by general human purposes accidentally bumping into each other. One might wonder whether this is just too fantastic. How could mutually beneficial arrangements suddenly appear without the parties intending such result? Game theorist Brian Skyrms presents some conceptual tools for understanding how this might be possible. Skyrms creates game-theoretic scenarios where the players detect the strategies of other players and is better off adopting these strategies if others do so. Eventually, the strategies of all participants come to correlate in what Skyrms calls a "correlated equilibrium" (Skyrms 1996, ch. 4). Imagine, for example, an intersection without any traffic lights where two cars meet at the same time but going in different direction. One driver has the other on her right-hand side, while the other on her left. By following the US rule that the driver on the right goes first, the two drivers act in a way that accords with a correlated equilibrium. This rule is, of course, taught at driving schools but suppose for a moment that they aren't. If the drivers start behaving in accordance with the rule, others will

[21] Cf. Dawkins' Orthodox Neo-Darwinist doctrine.

do so as well, because they are all better off following the rule if everybody else does it – they create a correlated equilibrium. The fact that the driver on the right goes first is completely random, it might as well be the driver on the left that went first. We can imagine that such correlated equilibriums can, and have been, established throughout populations where the participants learn and acquire beliefs about what strategies others are utilizing. Such a story doesn't need to postulate any hypothetical contract, but only the far more realistic picture of people who adjust their actions to that of other's in order to produce mutually beneficial results.

3.2 Taking stock: The problem of normativity

With the alternative genealogy developed in the last section in hand, we have a causal story of the development of the moral judgment that doesn't strictly appeal to genetic selection like on Joyce's account. Rather, I have argued that its more plausible that morality developed from cultural selection. Remember the formal evolutionary debunking argument presented in the introduction. The argument had a causal and one epistemic premise. For Joyce, moral nativism was the causal premise. The shift from evolution by genetic selection to evolution by cultural (group) selection amounts to a shifting the causal premise in the debunker's argument. As we shall see, this shift does not carry with it a debunking effect.

On the alternative view, moral cognition is the joint achievement of innate mechanisms and cultural information, where the explicitly *moral* structure stems from cultural information in the exogenous environment. Remember that culture is defined here as mental states that is acquired or modified by social learning and affects behavior. Since culture isn't, at least in any obvious way, determined or caused directly by genetic selection, Joyce's evolutionary debunking argument consequently fails. But as I briefly noted in section 2.4, Joyce's skepticism isn't only intended to be supported by moral nativism. He also offers a conceptual analysis of the moral judgment and claims that it involves properties that cannot be obtained. His account of moral nativism is also supposed to be a positive case for explaining why the moral judgment seems to have this illusory quality. It's supposed to work as an *error* theory. Because I share his analysis of the moral judgement, coupled with my rejection of moral nativism, it seems that we need a *success* theory – a theory that shows how those distinctive features of morality might be obtained. But what are those distinctive features? In chapter one I argued that there is something distinctive about morality that stand in need of explanation.

Moral judgments are, for example, different from prudential and conventional judgments. In section 1.5 I discussed the way that moral judgments are *inescapable*. This means that they apply to everyone regardless whether someone's desires or inclinations would be served by following that judgment. However, conventional judgments such as rules of etiquette also enjoy similar scope of application. Nonetheless, I argued, a simple empirical observation is that moral judgments usually override other kinds of judgments. They are thought to have a special kind of authority, which together with its inescapability, I called *moral normativity* (Cf. Brink 1997). All attempts at vindicating morality will need to properly account for moral normativity and an evolutionary approach to vindication must show how evolution played a part *without inevitably debunking morality*. The rest of the thesis will try to show how the cultural evolution of morality not only neutralizes the debunker by refuting nativism, but also forms a base for a vindicating account of moral normativity.

The puzzle I will try to answer in the following is typically referred to as *the problem of normativity*. In a nutshell, the problem can be posed in terms of the following question: how is it possible to recognize and act on a duty which applies and has overriding authority over you even though you have no motivation or inclinations to follow that duty whatsoever? Answering this question requires us to enter a discussion about practical reason – the apparatus of reasoning about what to do. The problem is usually targeted against naturalistic theories of morality because of the instrumental conception of practical reason that usually accompanies them. On the instrumental conception, practical reason is a purely formal faculty for guiding transitions from basic motivation to non-basic motivation and ultimately to action. Therefore, normative demands on the agent are contingent on there being some motivation that the agent already has. The apparent inescapability of moral demands, then, immediately creates an explanatory problem for evolutionary approaches and the instrumental conception of practical reason. At the core, the problem is that duties imbued with moral normativity seems to imply some objective moral realm because they work irrespective of subjective preferences. But an evolutionary approach to morality doesn't seem to allow for a moral realm corresponding to some sui generis portion of the world. As we have seen, Joyce's evolutionary debunking argument is targeted against the possibility of there being such a realm. If morality came about in the course of evolution by genetic selection for the reason of ensuring social cooperation and not for tracking moral facts or truths, that implies that there aren't such facts or truths to be found. Even if we refute moral nativism and approach things in terms of evolution by cultural group selection, as I have done, the problem doesn't

disappear. For on the alternative genealogy I presented in the former section, morality is in some sense constructed cultural information. There isn't a chunk of the universe in which the moral realm resides.

Recall our discussion in section 1.5. Here I argued that one way of thinking about the inescapability of moral demands is to think of them in the same way as we think of rules of etiquette. They are both nonhypothetical; rules that apply to you even though you have no motivation or inclinations to conform to them. But there is a difference between the practical authority morality and rules of etiquette. Kant thought that the inescapability of moral judgments implied that they are a binding law of practical reason and that makes their practical authority. However, since rules of etiquette enjoy similar scope of application we can see that this is wrong because no one supposes that rules of etiquette are a binding law of practical reason. On the instrumentalist conception of practical reason, as contrary to the Kantian conception, this suggests that even though moral demands apply to me, they don't necessarily provide me with reasons to conform to them unless I have some relevant existing motivation. Joyce (2006, p. 60-61) is unsatisfied with this conclusion. He shares Kant's intuition that there is something deeper about the normative force of moral judgments. He thinks it is very strange to grant that a moral demand applies to someone but does not necessarily supply that someone with reasons to satisfy it. Again, compare with etiquette. Joyce points out that there are reasons to behave politely independent of the agent's motivations. These are reasons provided by the *institution* of etiquette, such that given its rules, everyone has a reason to behave politely. Therefore, *morality as an institution* could also provide reasons independent of existing motivations. However, we normally think of moral reasons to have a practical authority that reasons provided by the rules of etiquette simply does not supply. A simple observation is enough to show this. For example, someone who does not care at all for the institution of etiquette can simply disregard the reasons it offers. We can imagine an anarchist who thinks that table manners are a sham and there is no point of conforming to the reasons they supply. We might think that this is unfortunate and impolite, but in the end, it's not the end of the world. If, on the other hand, for no reason, someone started hitting and kicking an innocent table partner, that would be unacceptable in a more serious way, because that person is now not conforming to the rules of morality. Therefore, as contrary to rules of etiquette, morality seem genuinely institution transcending. Moral judgment seems to yield genuine practical deliberative considerations for those ascribed by them.

That morality has this practical authority is an empirical claim and should not be taken as an a priori conceptual clam about moral judgment holding for all possible worlds. But if we aren't prepared to attribute this feature to moral judgments, I cannot see what makes morality different from conventional judgments such as those pertaining to the rules of etiquette. In one sense, we have "evidence" for the existence of morality in the robust findings that the moral/conventional distinction is a marked feature of our psychology. That is, there are evidence that we do treat the reasons provided by morality different than the ones implied by convention. For example, one important finding from the research on the moral/conventional distinction is our ability to treat moral judgment as not dependent on the will of the judger; that we treat them as authority-independent. This parallels the thought that morality is institution-transcending; morality isn't dependent on the authority of morality as an institution. How is it possible to make sense of this from an evolutionary approach to morality? One option might be Joyce's debunking account. For Joyce, morality only *purport* to have a normative grip on the agent and thus only *seem* to be institution-transcending. Since his evolutionary approach indicates an instrumental conception of practical reason, he cannot see how nonhypothetical reasons could ever transcend institutions and the existing motivations of agents. But he thinks we nonetheless do presuppose a transcendent normativity when making moral demands on each other. However, that doesn't mean that it follows that an external realm of moral reasons actually exists. Joyce thinks that there is no nonhypothetical reasons external to those specified by human institutions or the existing motivations of individuals. There must be some other explanation, then, and Joyce is in a position to provide one; although we make the fundamental error of presupposing a transcendent normativity that doesn't exist, the error is adaptive. According to Joyce's hypothesis, the innate moral sense has the function of generating a cognitively rich emotion of guilt when failing to cooperate. To avoid strong weakness of the will-type problems, then, its expedient for the moral sense to presuppose a concept of moral reasons that transcend institutions and apply independently of the existing motivations of the agent. Thus, the extra 'oomph' provided by the moral sense (or conscience; Joyce uses conscience and moral sense somewhat interchangeably) is Joyce's explanation for the apparent practical authority of moral reasons. Put another way: moral nativism *debunks* the distinctive normative ground upon which we think morality lie. On Joyce's account, the *descriptive* side of evolutionary ethics explains away the room for a normative counterpart. It provides us with an explanation for why morality genuinely *seems* to be imbued with a special authority but doesn't grant that authority to be normatively real. However, an evolutionary vindication, which I am pursuing, provides an explanation for why and how that authority is normatively real but equally constrained in terms

of evolutionary explanations as the nativist account. On this approach, instead of debunking the existence of moral reasons, the descriptive side of the explanation leave room their existence and accounts for role played by evolution in the construction of those reasons.

3.3 What is a moral reason?

Regarding moral normativity, I agree with Joyce on two important matters. First, we do treat moral judgements as not contingent on the satisfaction of some existing motivation of the agent. Second, we also treat moral judgments as providing reasons for action. It seems very strange indeed to say that I ought not to steal, but I have *no reason* to do so. Moral 'oughts' implies reasons for those subject to the prescription. But the crucial question for us here is: what kind of reasons are implied by moral prescriptions? The reason we are looking for is usually called a *justifying reason*. From the moral perspective, to say that I ought not to steal, is to say that there is something about my situation that justifies me to stop stealing: for example, that I would benefit from depriving others of their earned property. We could say that the norms concerning stealing is part of morality as an institution, and the institution is the source of this particular justifying reason. But, as we have seen, moral reasons must involve something more. Otherwise, reasons that justifies action from the perspective of conventional institutions, such as the rules of etiquette, could have equal normative grip on its subjects. The missing link, I contend, is what we call a *motivating reason*. A motivating reason is a consideration that hooks up with the agent's motivational system such that the recognition of that reason by the agent generates motivation to follow through on that action. Motivating reasons are different from justifying reasons. For example, think again on weakness of the will-type problems. We may have strong innate dispositions to produce motivations that go against cooperation – that is, we have strong selfish dispositions – and there are good chances that the motivating reasons we recognize to favor selfish action will defeat the justifying reasons acting out of concern for others. But it can also go the other way around. There are some justifying reasons that by itself is motivating and have the power of overriding existing motivating reasons. A moral reason, then, is a reason that is *both a justifying reason and a motivating reason*.

We are now in a position to clearly describe the problem facing the evolutionary vindication of morality currently under review. Joyce cannot see how a justifying reason could be a motivating reason. To see why, we have to introduce two theoretical options concerning the origin or source of reasons; internalism and externalism about reasons. Internalism about reasons is the

claim that the reason must be embedded internally in the individual's motivational system. On the other hand, recognizing that one has a reason for action seems to involve the recognition that there is something of value that would be realized by doing it; that there is a consideration that weighs in favor of doing it. One could fail to recognize such a consideration, and that would imply that the reason was there independently of the agent. Therefore, externalists about reasons claim that reasons are external to the motivational systems of individuals. While subscribing to an instrumental theory of practical rationality, Joyce also endorse the closely related view of internalism about reasons. Together they can explain *how* an agent could act on moral reasons; the formal faculty of instrumental reason recognize a reason embedded in the motivational system of the agent and produces the necessary motivation for following through on that action. However, as we have described a justifying reason above, for example, a situational feature concerning the institution of morality or etiquette, they clearly seem to be external to the motivational system of the agent. On Joyce's account, then, how can we know that they are moral reasons, if the information is not outside the motivational systems of the agent? His explanation, as we have already seen, is that this information is innately specified. The problem is that concerning justifying reasons, externalism seems most plausible, but for motivating reasons, internalism is most plausible. A system of norms supplying justifying reasons are clearly independent of any individual, but it's difficult to see how they can become motivating if there aren't any existing inclinations that would be served by recognizing that reason. Nonetheless, we observe that moral reasons have the feature of being both justifying and motivating.

To solve this impasse, I will utilize a constructivist conception of moral reasons developed by David Wong (see ch. 7 2006; 2008; 2009). According to his conception, moral reasons are not objective in the realist sense – that is, they don't reside in some sui generis portion of the universe. Rather, moral reasons are objective in the sense that they are *inter-subjective*. Put in the relevant terms, moral reasons are *external* to a particular individual's psychology (or motivational system), but *internal* to human psychology (or motivational propensities) in general. For my purposes, the explanatory advantage of Wong's conception of moral reasons is the role played by socialization in shaping our general human motivational propensities – that is, our relevant innate capacities – and attaching them onto morally appropriate objects. Basically, this boils down to a description of how the moral judgment, understood as the capacity to recognize moral reasons, is thought and learned through socialization. An important distinction here is that between the motivational forces of *propensities* and those of *desires*. A

desire is often understood as having fixed and determinate intentional objects, which is situationally stable and invariant. A propensity, however, can be understood as being much more flexible and often is dependent on the situation or context for determining its intentional objects. Propensities is better able to explain the indeterminate motivational forces relevant to moral behavior. Just consider how our propensity to help others in distress can vary according many different situational factors such as the degree to which the person(s) in distress resemble the agent or if they appear young and vulnerable. Another example of this is the way human psychology have a bias toward in-group solidarity but is significantly flexible in choosing its group. Therefore, Wong (2009, p. 344) writes:

> "*Learning the moral reasons that go beyond such factors, and subsequently recognizing them on different, varied occasions, can render more stable the propensity to help over a variety of situations and hence can make it a more determinate helping response to need or distress. But reasons could not play this role if having a moral reason were dependent on already having such a response, if they were, as Bernard Williams argued, "internal" to the "motivational sets" of individuals. Moral reasons are internal to the propensities, not of each individual who has these reasons, but to the propensities most human beings generally have and to the propensities they could be brought to have through socialization. The content of moral reasons and the demands they make of agents are limited by human psychology in general, even if they are not limited by the particular psychologies of individuals. In that sense, moral reasons are internal to human psychology.* "

On this conception, moral reasons serve a critical function – it shapes and structures our domain diverse innate emotional systems. By learning moral reasons individuals gets socialized into being stable cooperative partners. As the reader might already have figured out, this conception of moral reasons fits nicely with the alternative genealogy of morality presented above. Accordingly, my proposal is to view moral reasons as cultural artifacts constructed in the course of cultural evolution. Cultural information, as defined here, can take the form of cues in the socio-cultural environment and is therefore external to the psychology of the individual but internal to human psychology in general. Moral reasons, therefore, gets their practical authority from the way that our motivational propensities are shaped by socio-cultural cues in the environment. Keep in mind that the hypothesis under consideration is that this has occurred for thousands of years and that the most successful groups have been culturally selected for in the

course of evolution. This hypothesis critically depends on the possibility of how the learning and recognizing of moral reasons can shape our motivational propensities – i.e. *how a justifying reason can become motivating*. In line with the general methodological framework I have been pursuing, I will treat this as an empirical question. I have argued here that moral cognition is the result of a sophisticated relationship between emotion and cognition. I pointed out in section 1.4 that, prima facie, the moral judgment as a speech act can both be expressing cognitive and non-cognitive content. In section 1.6, I also presented some evidence that our moral emotions such as guilt is cognitively rich. In the following section, I follow up on this theme to present a case for how the cognitive achievement of recognizing moral reasons, might affect our emotional motivational systems. This might sound Kantian and cognitivist, but as we will see, coupled with the evolutionary picture of the relationship between emotion and cognition, the hypothesis undermines the stark dichotomy between cognitivism and non-cognitivism.

3.4 The practical authority of moral reasons: how a justifying reason can become motivating

To sum up some of the things that have been argued in these previous sections. What is a reason? Reasons are those considerations weighing in favor of or against an agent's doing something. Following Wong's (2008, p. 248) apt description of how they are structured, we can say that they are:

> *"…three-place relations between an agent A, an action X, and a feature F in the agent's situation that weighs in favor of A's doing X. For example, A may have a reason to help B in virtue of B's being in imminent danger of being harmed and A's being able to help with no risk and low personal cost to herself".*

Defined as such, we call this a justifying reason: F is the feature that purport to justify A's doing X. However, as we have seen, a justifying reason is not necessarily a motivating reason. It might not motivate A to do X. Moral reasons, that is, reasons with a special kind of practical authority, is a justifying reason that can become motivating. The main purpose of this section is to show how this could be possible. However, as I have indicated above, the evolutionary relationship between moral reasons and motivational propensities is crucial to this explanation, so I will start there.

I understand a motivational propensity to be a type of functional state in which dispositions to act or feel is grounded (Wong 2008, p. 251). The thirsting for a drink of cool water, the crave for a particular food, sexual arousal, and more importantly for us here: the strong disposition to help a person in distress are all examples of such states, where there is a feeling of *urge* toward an intentional object. Human motivational propensities are rooted in human nature as a part of our bioprogram, naturally selected for in response to various selective pressures in the course of evolution. Consider the multitude of motivational propensities that could be said to be innate to human beings. As I argued in section 2.4, there is great need to shape and structure our motivational propensities in order to make us stable cooperative partners. Cultural norms, I argued, played a special role in regulating ourselves in this way conferring an evolutionary advantage on those successfully implementing culture within their group (cf. Richerson and Boyd 2005). I suggest that one way in which culture was successfully implemented within a group was by *embedding reasons in the existing motivational propensities of their members*. Remember that reasons are those features of situations that go into the identification of what to do. Cultural norms evolved as new members of a group imitated the solutions to problems previously worked out by the most successful members of the group. Cultural norms, then, could be described as specifying which situational features (i.e. which reasons) counts as the appropriate considerations for what do to. In this way cultural norms shapes human motivational propensities to identify the appropriate intentional object by embedding reasons within them. On my hypothesis, in response to the selective pressure to ensure social cooperation, moral norms culturally evolved through cultural group selection. With the development of moral norms came the embedding of *moral* reasons within our propensities. In this sense, then, moral reasons get culturally constructed to perform a certain function. Therefore, they are external cultural cues, but they are internal to human psychology in general in the sense that they are constrained by the propensities that human beings can possible have.

The view I am proposing, then, is that the way a justifying reason becomes motivating is by embedding them in our evolved motivational propensities. This is done by designating certain features of situations as considerations for a certain action. The nature of such information is socio-cultural, and as I have argued, our brains are wired to handle such information. But are they wired to handle them in the correct way? The view presupposes two premises about evolutionary psychology that needs to be made explicit: 1) that our evolved motivational systems is sufficiently imbued with cognitive sophistication, and 2) that they are sufficiently

malleable. 1) because we need to be able to recognize different situational features as reasons for acting towards some intentional object. For example, to be able to care for kin, we need to be able to recognize features of other individuals that indicate kin, and such recognition gives us cues – or might we say *a reason* – to act altruistically towards that individual. But our motivational propensities are not desires, they are significantly general. Therefore, 2) indicates that they can be "trained" to recognize cues that are different from the original situational feature generating motivation. That is, the original situational feature in our ancestor's environment creating a selective pressure for a fitness enhancing response. It follows from 2), however, that a third premise must also be made: 3) that we can be thought which situational features that should generate motivation to act, and the nature of that motivation. Premise 1) is uncontroversial: we need to be able to recognize certain features in our environment for basic reactive responses. For example, that we have distinct responses to what we recognize as danger is enough to settle this premise. Premise 2) and 3), however, needs further justification.

In taking the view, 2), that our motivational systems are flexible to a significant degree, might at first sound like practical rationality is significantly autonomous from our evolved motivational propensities. This would not sit well with the hypothesis I am currently defending, because it would potentially mean that I could not connect the recognition of reasons with the motivational efficacy of those propensities. However, I argue that the picture of evolutionary psychology currently emerging from cognitive science show how the issue is more complex than assumed by the autonomous reason view. The model I am referring to is popularly known as the dual process theory (see Daniel Kahneman 2011 for an accessible and popular overview). Dual process theory orders the way human beings process and react to information in their environment into two separate tracks. One fast and automatic track, and one slow and deliberative track. The crucial thing for my purposes is that both tracks can be involved in the complex process of having an emotion. It is the fast and automatic track, however, that should undermine our confidence that practical rationality is significantly autonomous from our evolved emotion-based motivational propensities. In section 1.6, I discussed the way that our basic emotions are adaptive mechanisms, selected for by biological natural selection to do a certain job involving psychological and physiological elements triggered by environmental factors, for example, factors such as facial expressions (Cf. Ekman & Robinson 1994). Very often, these mechanisms work independently of each other, each triggered by specific types of stimuli and ignorant of stimuli that is not of the right sort – that is why they are very hard to "reason" with. I called on the example of the autonomous nervous system; in response to

potential threats, the autonomous nervous system has processed information and started relevant processes (fight, flight or freeze) long before you are fully conscious of what's going on. It's not only the autonomous nervous system that works in this way, the dual process theory generalizes that emotions very often involves this fast and automatic processing. Recall psychologist Paul Rozin's studies on the emotion of disgust, where subjects had powerful disgusting reaction sterile plastic dog feces and perfectly edible fudge that looked like poop. Even though the subject *knew* that those objects are not hazardous to their health in any way, the fast and automatic processing of the disgusting emotions recognized the situation differently.

But there are avenues for revising the initial fast and automatic emotional response on a slower, deliberate and conscious mode of processing. For example, if detecting a danger in the environment, it can take the form of evaluation of the specific degree or the way the object is to be feared. It might involve more sophisticated types of reflection such as wondering why one is feeling this way or what it is out there that is causing one to react in that particular way. Connected to the fast track in this way, the slower track can result in choosing different modes of action or at least a modification of that mode of action, but still tightly linked with the initial fast response. The modified mode of action, then, involved some conscious reflection but retained the motivational efficacy of the initial fast response. Jonathan Haidt's view of moral judgment, which I mentioned in section 2.1, could also be an example of the two-track model. On his view, the moral judgment is a had hoc justification of an initial fast and automatic emotional response. Moreover, although Rozin's experiments show that the initial fast response is powerful and hard to override, we could perhaps expect his subjects to eventually subdue their reaction if finding themselves in the given situation more frequently. Hypothetically, after experiencing similar situations over and over, in the end one should perceive enough regularities as to figure out which response produce the optimal and appropriate reaction. Clearly, recognizing that the apparent disgusting objects really are fakes would provide the subjects with *sufficient reason* not to display powerful disgusting reactions. But the fact that human beings must experience that what one initially thought was disgusting or dangerous over and over again before acting on such reasons, shows that there is no default linkage between action and judgment. In fact, in light of the two-track model, the default seems rather to be that we feel and act against what we know there to be sufficient reason to do. This claim might be too strong, however, the dual process theory does suggest that the fast and automatic processing is more ancient than the slow and deliberate mode and that might explain why the fast and automatic track seem to be in the driving seat.

It's not surprising that human beings are prone to act against what one has sufficient reason to do when considering that practical rationality might have been created on top of the oldest layers of our motivational systems. If this is true, then our apparatus for practical reasoning is dependent on and therefore easily overridden by these ancient layers for their motivational efficacy. One response to this picture of our ability for autonomous reason is pessimism about rationality. In one sense, this is understandable, but the correct response in my view, is to realize that piggybacking on these old layers is what makes it possible to act on reasons in the first place. Rationality might not be the default, but at least it can be a 'precarious achievement' to invoke Wong's optimistic phrase. For example, fear of dangerous animals such as large predators are plausibly rooted in the oldest layers of our motivational systems. The point here is that if our modern fears, such as fear of financial crises are to be possible, there must already be in place relevant neural circuits in the brain for new fears to piggyback on for motivational efficacy. As Wong (2009, p. 354) writes:

> "*Human beings may be wired to respond to threats to their physical wellbeing, but the response may get attached to an expanded category of threats through associative links formed in experience, learning, and higher cognitive functioning. Cultural practices and institutions help to enormously expand what constitutes a fear".*

Evolution is a significantly conservative process and "prefers" to build on top of existing mechanisms. Instead of creating novel systems for solving problems, evolution goes to work on what is already there, originally designed for some other purposes. The result might be messy and clumsy, but one that is adaptive enough. In establishing 2), then, it must be acknowledged that our motivational propensities are not fundamentally plastic, but that reasons can gain foothold by further developing their original function. How such reasons can gain this foothold will ultimately depend on establishing premise 3): how we learn these reasons.

The hypothesis under consideration is that our capacity for moral judgment as part of practical rationality piggybacks for their motivational efficacy on older systems that humans have evolved. I started out describing the problem of normativity by considering the apparent failure of the instrumental conception of rationality for showing how we can go from normative premises to action. It holds that we can only rationally identify means to ends provided with desire. Therefore, to be motivated for a particular action, one must recognize that the reasons

on the offer will satisfy some desire that one already has. When it comes to moral motivation, then, the proponents of the instrumentalist conception run into some problems. They postulate that the agent has some desire such as "sympathy" or desires to relieve the distress of others that approximate the objects of morality (Foot 1972). But as we have seen before, these belong to capacities that are not strictly moral. The solution I suggest lies in an understanding of how desires that are not moral in content, can come to have moral content. Using the same piggyback hypothesis, I propose that learning to recognize moral reasons can shape an individual's motivational propensities. The idea has been present throughout the thesis: that capacities such as psychological altruism and the norm acquisition device are innate precursors that are necessary for our capacity for moral judgment. But they are not sufficient – the crucial *moral* structure comes from cues in the socio-cultural environment. Accounting for 3) therefore depends on how we use our existing prosocial capacities to develop them into moral capacities.

One influential theory of moral development, due to psychologist Martin Hoffman, utilizes this picture of morality and suggests how young children start to internalize morality in their early encounter with empathetic distress when witnessing the distress of another. Not surprisingly, our innate capacity for imitation is relevant here. Imitative learning was, as Tomasello (1999) has stressed, crucial for the evolution of culture, and if I am correct in that morality is a cultural artifact, we should expect that moral learning uses the same learning mechanisms as culture. Hoffman describes how young children imitate the facial expressions, bodily postures and tone of voice of those in distress, before they themselves start showing signs of the of distress by experiencing the responses brought about by imitation (Hoffman 2000). Hoffman invoke the idea of an "induction" to show how we get socialized and taught which situational features counts as moral considerations. Imagine that a child has pushed down another child causing her distress. An adult may then express disapproval and contempt by pointing out the consequences of her action (induction). The adult might say something like "Bad! That made her cry" or "That made her feel unhappy". According to the account of moral reasons given here, then, these are the examples of the early springs of moral reasons in the process of moral socialization. The justifying reason of not pushing down the other child becomes motivating because in calling attention to the distress of the victim, the adult hooks the induction up with the child's empathic proclivities to ultimately generate empathic distress. The induction would also point out how it was the actions of the child that lead to the distress of the victim – how it was her *fault*. In turn this leads to the conditions for feeling guilt. To relieve the child from such emphatic distress and guilt, the adult may suggest that the child could perform reparative acts such as apologizing

for her actions. Hoffman supposes that when such chain of events – wrongdoing, induction, empathic distress and apology – is repeated over and over, it taps into the child's memory code to clout future decisions and behavior. This is the way in which moral reasons get embedded in our proclivities and which ultimately accounts for the practical authority of moral judgments.

Conclusion:

On the competing story I have considered, moral normativity is only *experienced* as genuine. It only seems to be real because built into our human nature there is an illusion. An illusion of a transcendent moral realm of which our practical reasoning about moral reasons purport to grasp facts or truths. Evolved by genetic selection, this illusion serves the function of generating fitness increasing behavior and that's enough for explaining its emergence. If this is true, there is no need to appeal to an external moral realm to explain manifest morality. In fact, if moral nativism is true, its simply very unlikely and implausible that our moral judgments have the ability to track facts, simply because there can't be any such facts to track. It's not as if there was an existing moral realm of which the moral judgment evolved to track. On Joyce's account, the *moral* came into being when the correct genotype started to spread and that's all there is to the story. Therefore, the debunking effect is that moral normativity isn't genuine. However, this whole story depends on the truth of moral nativism, and as we have seen, there are better options.

In the introduction I laid out that an evolutionary vindication of morality would have to be able to justify morality from within *the moral perspective*. I claimed that this is different from instrumental justifications of morality where morality is only justified as useful behavior from the outside. Compare with religion. Religion cannot be properly vindicated in the relevant sense without a justification for the existence of God. Accordingly, morality cannot be properly vindicated without a justification for the existence of moral reasons. I have presented a theory of how moral reasons do exist as cultural artifacts. On this view, in one sense there is a moral realm, only that this realm isn't a sui generis part of the universe. Rather, the realm consists of cultural information defined as mental states that is acquired or modified by social learning and affects behavior. Cultural norms and eventually moral norms are the result of thousands of years of the social embedding of those norms. I contend that there is a progressive story to be told as well and I want to end by briefly show how this account give rise to a type of 'normative ethics' based on the idea that morality resulted from this social embedding.

Recall that the epistemic premise in the evolutionary debunking argument is that evolution by natural selection is an off-track process. But, this is only true in the case of morality if by natural selection we mean genetic selection. I mentioned in the introduction that altering the causal premise, as I have done here, would have consequences for the epistemic premise. If we change

the causal premise from evolution by genetic selection to evolution by cultural selection, what becomes of the epistemic premise? Is evolution still an off-track process? It seems to me that the reason why some groups outcompeted others in in response to cultural group selection, was due to them being more successful in meeting the selective pressures. Those pressures were identified as social cooperation. Thus, we can say that those were the groups who most successfully acted according to facts about social cooperation. This, I think, opens up for the thought that, in some sense, morality developed as a reliable way to track facts about social cooperation. And those groups who cultivated morality, those who sat around the campfire deliberating and putting forward moral reasons in defense for their solutions tracked those facts the best. By socially embedding morality and in this way, what they essentially did inventing normative ethics: the cultivation of morality.

References

Acharya, S., & Shukla, S. (2012). Mirror Neurons: Enigma of the Metaphysical Modular Brain. *Journal of Natural Science, Biology and Medicine*, 118-124.

Altham, J. E. (1986). The Legacy of Emotivism. In G. Maconal, & C. Wright, *Fact, Sience and Morality* (pp. 275-288). Oxford: Basil Blackwell.

Axelrod, R. (1984). *The Evolution of Cooperation.* Baisc Books.

Blair, R. J. (1995). A cognitive developmental approach to morality: Investigating the psychopath. *Cognition*, 1–29.

Blair, R. J., Jones, L., Clark, F., & Smith, M. (1997). The psychopathic individual: A lack of responsiveness to distress cues? *Psychophysiology*, 192–198.

Boehm, C. (1999). *Hierarchy in the Forest: The Evolution of Egalitarian Behavior.* Cambridge: Harvard University Press.

Boyd, R., & Richerson, P. (1985). *Culture and the Evolutionary Process.* Chicago: Chicago University Press.

Boyd, R., & Richerson, P. (2005). *Not by Genes Alone: How Culture Transformed Human Evolution.* Chicago: Chicago University Press.

Brink, D. (1997). Kantian rationalism: Inescapability, authority, and supremacy. In G. Cullity, & B. Gaut, *Ethics and Practical Reason.* Oxford: Oxford University Press.

Carlsmith, K., Darley, J., & Robinson, P. (2002). Why Do We Punish? Deterrence and Just Deserts as Motives for Punishment. *Journal of Personality and Social Psychology*, 284-299.

Chandra, S. (2005). Punishment and the Strategic Structure of Moral Systems. *Biology & Philosophy*, 767–789.

Clyne, B. (2015). Nativism and The Evolutionary Debunking of Morality. *Review of Philosophy and Psychology*, 231-253.

Cowie, F. (1999). *What's Within?* Oxford: Oxford University Press.

Cummins, D. (1996). Evidence of deontic reasoning in 3- and 4- year old children. *Memory & Cognition*, 823–829.

Damasio, A. (1994). *Descartes' Error: Emotion, Reason and the Human Brain.* Quill.

Dawkins, R. (2006). *The Selfish Gene.* Oxford: Oxford University Press.

de Waal, F. (1989). *Peacemaking Among Primates.* Cambridge: Harvard University Press.

Dennett, D. (1995). *Darwin's Dangerous Idea.* Simon and Schuster.

Dewey, J., & Tufts, J. (1932). *Ethics, 2nd ed.* . New York: Henry Holt.

Dwyer, S. (1999). Moral Competence. In K. Murasugi, & R. Stainton, *Philosophy and Linguistics.* Westview Press.

Ekman, P., & Robinson, R. (1994). *The Nature of Emotion.* New York: Oxford University Press.

FitzPatrick, W. (2014). Debunking Evolutionary Debunking of Ethical Realism. *Philosophical Studies*.

Flanagan, O. (1991). *Varieties of Moral Personality: Ethics and Psychological Realism.* Cambridge: Harvard University Press.

Flanagan, O. (2016). *Geography of Morals: Varieties of Moral Possibility.* Oxford: Oxford University Press.

Fodor, J. (1983). *The Modularity of Mind.* Oxford: Oxford University Press.

Foot, P. (1972). Morality As a System Of Hypothetical Imperatives. *The Philosophical Review*, 305-316.

Foot, P. (1995). Does Moral Subjectivism Rest On a Mistake? *Oxford Journal of Legal Studies*, 1-14.

Ghiselin, M. (n.d.). *The Economy of Nature and the Evolution of Sex.* Berkely: University of California Press.

Gould, S. J., & Lewontin, C. (1979). The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 581-598.

Greene, J. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, 1144-1154.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, 517–523.

Griffiths, P. (1997). *What Emotions Really Are: The Problem of Psychological Categories.* Chicago: University of Chicago.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment.". *Psychological Review*, 814–834.

Haidt, J., & Joseph, C. (2004). Intuitive ethics: how innately prepared intuitions generate culturally variable virtues. *Daedalus*, 55-66.

Hamilton, W. D. (1964). Genetical Evolution of Social Behavior 1 & 2. *Journal of Theoretical Biology* , 1-52.

Hare, R. (1993). *Without Conscience. The disturbing world of the psychopaths among us.* The Guilford Press.

Hauser, M. (2006). *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong.* New York: Harper Collins.

Hoffman, M. (2000). *Empathy and Moral Development.* Cambridge: Cambridge University Press.

James, S. E. (2010). *An Introduction To Evolutionary Ethics.* Wiley-Blackwell.

Joyce, R. (2006). *The Evolution of Morality.* Cambridge: MIT Press.

Joyce, R. (2007). Is Human Morality Innate? In P. Carruthers, S. Laurence, & S. Stich, *The Innate Mind: Volume 2: Culture and Cognition.* Oxford: Oxford University Press.

Joyce, R. (2016). *Essays In Moral Skepticism.* Oxford: Oxford University Press.

Kahane, G. (2010). Evolutionary Debunking Arguments. *NOUS*, 103–125.

Kahneman, D. (2011). *Thinking Fast and Slow.* Farrar, Straus and Giroux.

Kitcher, P. (1998). Psychological Altruism, Evolutionary Origins and Moral Rules. *Philosophical Studies*, 283-316.

Kitcher, P. (2005). Biology and Ethics. In D. Copp, *The Oxford Handbook of Ethical Theory.* Oxford: Oxford University Press.

Kitcher, P. (2011). *The Ethical Project.* Cambridge: Harvard University Press.

Kitcher, P. (2012). Naturalistic Ethics Without Fallacies. In P. Kithcer, *Preludes to Pragmatism: Toward a Reconstruction of Philosophy* (pp. 303-324). Oxford University Press.

Kramer, P. (1993). *Listening To Prosac.* Penguin USA.

Laurence, S., & Margolis, E. (2001). The Poverty of The Stimulus Argument. *British Journal for the Philosophy of Science*, 217-276 .

Mackie, J. (1977). *Ethics: Inventing Right and Wrong.* Harmondsworth: Penguin.

Mellars, P. (1995). *The Neanderthal Legacy: An Archaeological Perspective from Western Europe.* Princeton University Press.

Mivart, G. J. ((1871) 2008). *On the Genesis of Species.* BiblioLife.

Nichols, S. (2002). On the genealogy of norms: A case for the role of emotion in cultural evolution. *Philosophy of Science*.

Nichols, S. (2004). *Sentimental Rules: On the Natural Foundations of Moral Judgment.* Oxford: Oxford University Press.

Nichols, S. (2005). Innatenes and Moral Psychology. In P. Carruthers, S. Laurence, & S. Stich, *The Innate Mind: Structure and Contents* (pp. 353-370). New York: Oxford University Press.

Nucci, L. (2001). *Education In The Moral Domain.* Cambridge University Press.

Nucci, L., & Weber, E. (1995). Social Interactions in the Home and the Development of Young Children's Conceptions of the Personal. *Child Development*, 1438-1452.

Prinz, J. (2007). Is Morality Innate? In W. Sinnott-Armstrong, *Moral Psychology.* Oxford: Oxford University Press.

Rosenberg, A. (2002). Darwinism in moral philosophy and social theory. In J. Hodge, & G. Radick, *The Cambridge Companion to Darwin* (pp. 310-332). Cambridge University Press.

Rozin, P., Haidt, J., & McCauley, C. (2000). Disgust. In M. Lewis, & J. Haviland, *Handbook of the emotions, the second edition* (pp. 637-653). New York: Guilford.

Ruse, M. (2009). Evolution and Ethics: The Sociobiological Approach . In M. Ruse, *Philosophy After Darwin.* Princeton University Press.

Samuels, R. (2002). Nativism In Cognitive Science. *Mind and Language*, 233-265.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision making in the Ultimatum Game. *Science*, 1755–1757.

Schmidt, E., & Bonelli, R. (2008). Sexuality in Huntington's disease. *Wiener Medizinische Wochenschrift*, 78–83.

Scott Curry, O., Mullins, D. A., & Whitehouse, H. (2019). Is It Good to Cooperate? Testing the Theory of Morality-as-Cooperation in 60 Societies. *Current Anthropology*, 47-69.

Shafer-Landau, W. (2012). Evolutionary Debunking, Moral Realism, and Moral Knowledge. *Journal of Ethics & Social Philosophy*, 1-37.

Singer, P. (2005). Ethics and Intuitions. *The Journal of Ethics*, 331-352.

Skyrms, B. (1996). *The Evolution of The Social Contract.* Cambridge University Press.

Smetana, J., Schlagman, N., & Walsh Adams, P. (1993). Preschool Children's Judgments about Hypothetical and Actual Transgressions. *Child Development*, 202-214.

Sober, E., & Wilson, D. S. (1998). *Unto Other: The Evolution and Psychology of Unselfish Behavior.* Cambridge: Harvard University Press.

Sripada, C. (2008). Nativism and moral psychology: Three models of the innate structure that shapes the contents of moral norms. *Moral Psychology*, 319–343.

Sripada, C., & Stich, S. (2007). A framework For The Psychology Of Norms. In P. Carruthers, S. Laurence, & S. Stich, *Innateness and the Structure of the Mind* (pp. 280-302). London: Oxford University Press.

Stich, S. (2008). Review: Some Questions about "The Evolution of Morality. *Philosophy and Phenomenological Research*, 228-236.

Street, S. (2006). A Darwinian Dilemma for Realist Theories of Value. *Philosophical Studie* , 109-166.

Tangney, J. (1992). Situational Determinants of Shame and Guilt In Young Adulthood. *Personality and Social Psychology Bulletin*, 199–206.

Tomasello, M. (1999). *The Cultural Origins of Human Cognition.* Cambridge: Harvard University Press.

Trivers, R. (1971). The Evolution of Reciprocal Altruism. *The Quarterly Review of Biology*, 35-57.

Turiel, E. (1983). *The Development of Social Knowledge: Morality and Convention.* Cambridge University Press.

Turnbull, C. (1971). *The Mountain People.* New York: Simon and Schuster.

Wastell, C., & Booth, A. (2003). Machiavellianism: An alexithymic perspective. *Journal of Social and Clinical Psychology*, 730-744.

Wong, D. (2006). *Natural Moralities.* New York: Oxford University Press.

Wong, D. (2008). Constructing Normative Objectivity In Ethics. *Social Philosophy and Policy*, 237-266.

Wong, D. (2009). Emotion and the Cognition of Reasons in Moral Motivation. *Philosophical Issues: Metaethics*, 343-367.