Joseph Paul Walsh

PhD Philosophy

Evolutionary Ethics without the Error: How Care Ethics Can Vindicate Moral Realism

88,915 words

Abstract

In this thesis I defend a form of moral realism against Richard Joyce's evolutionary argument for an error theory. I explain how evolutionary data can be used to explain human behaviour, ultimately endorsing a developmental systems perspective on the evolution of traits. I argue that evolutionary theories of ethics, developmentally conceived, are best demarcated from non-evolutionary ethical theories by appealing to the distinction between moral philosophy and moral psychology. I then set out Joyce's argument for an error theory, and in so doing respond to his claim that moral properties cannot be successfully naturalised. I then consider different naturalistic approaches to moral realism, assessing whether these approaches successfully meet Joyce's sceptical challenge. I look first at Philippa Foot's neo-Aristotelian approach to virtue ethics, arguing that her position fails because of her commitment to *eudaimonism*, and to a welfarist conception of function. I then consider Jesse Prinz's realist sentimentalism. This too, I argue, fails to constitute a convincing reply to Joyce, owing to internal inconsistencies, and to the failure of Prinz's theory to meet certain criteria intuitively constitutive of moral realism. Finally, I argue that a successful realist response to Joyce can be made by developing an evolutionary account of care ethics. I begin to develop such an account in the final chapter of the thesis, showing how the theory which I sketch meets each of the aspects of Joyce's argument for an error theory.

Contents

Introduction to the Thesis

Darwin changed everything. Not all at once, of course, and by no means inevitably. His inability to explain how offspring inherited traits from their parents meant that, only twenty years after its publication, Darwin's theory of natural selection "was thought to be on its deathbed".[1] If Mendel's research into the transmission of traits in peas had not been rediscovered in 1900, then the mechanics of Darwinian inheritance would have remained a mystery, and Darwin's legacy might have been quite different.[2] As it happens, though, Darwin changed everything. By showing that a-teleological, mechanistic selection pressures were capable of generating speciation, and that the human animal was no less likely to be a product of their operation than were non-human life-forms, Darwin's work challenged the traditional conception of humanity's place in the universe. The repercussions of this challenge are still being felt today, with an increasing number of human behaviours being studied from an evolutionary perspective as the possible products of natural selection. Evolutionary analyses of human behaviour promise to shed important new light on why we think and act as we do, and have been applied to areas of research as diverse as economics,[3] criminology,[4] and literary theory.[5]

Evolutionary theory has also been applied to the study of morality, and has been fruitfully used to suggest how natural selection could have favoured organisms well-disposed towards cooperation and reciprocal acts of kindness. However, according to some philosophers, such as Michael Ruse and Richard Joyce, this is not all that evolutionary theory has to tell us about morality. The influence which evolution has had on the formation of our moral beliefs, they argue, shows us that moral discourse rests on a fundamental mistake. That is, they argue that an evolutionary perspective on morality leads to moral error theory.

According to moral error-theorists, when we make the moral judgement that, for example, rape is morally wrong, or that we are morally obliged to give money to charity, we are making a statement of putative fact; we take ourselves to be saying something true, or at least capable of being true. That is to say, we are moral realists. The error theorist also claims, however, that in reality there *are* no moral facts: nothing is morally good, and nothing is morally bad. No moral obligations exist. Because ordinary moral discourse presupposes the existence of such facts, the error theorist concludes, it is systematically in error.

Moral error theories are, or at least ought to be, disquieting. They purport to challenge the coherence of everything we say and believe about morality, and they call upon us to radically revise the way in which we use moral language. Moreover, error theories cannot be summarily dismissed by emphatically insisting on the *obviousness* of, for example, murder's immorality; after all, this is exactly what the error theorist would *expect* someone in the grip of an error to say. Indeed, many of the error theorist's arguments are philosophically compelling. These arguments have persuaded

---

[1] Jablonka and Lamb (2005) p. 21.
[2] See Jablonka and Lamb (2005) Chapter 1 for more detailed discussion.
[3] See Aldrich et al (2008).
[4] See Cohen and Machalek (1988).
[5] See Carroll (1995).

some philosophers[6] that the best means of guarding against being convicted of an error is to deny that ordinary moral discourse is in the business of making factual claims.[7] Rather, it is argued, moral discourse is used to express the speaker's attitude towards certain actions. Because expressing an attitude is not the same as making a putatively factual, descriptive claim, the alleged error is thereby neatly excised from moral discourse. Ethical debate is then free to carry on as it always has. However, giving up on realism is a high price to pay for the conceptual coherence of moral discourse. Furthermore, if the error theorist's scepticism about moral facts is misguided, paying this price can be altogether avoided. Moral discourse can be shown to be coherent without embracing expressivism.

In this thesis, I develop a realist reply to Joyce's argument for an error theory. Joyce's theory has been highly influential, not least because it is developed from fairly uncontroversial evolutionary data. As Joyce presents that data, the facts about human evolution seem unequivocally to point towards an error-theoretic interpretation of moral discourse.

My reply to Joyce will be quite conciliatory, as I agree with much of what he has to say about morality. Thus, I do not deny that evolutionary forces have influenced the development of human morality. Nor do I reject Joyce's claim that ordinary moral discourse has realist aspirations. I part company from Joyce, however, in that I do not find a commitment to moral realism to be metaphysically problematic. It is possible to make the sorts of claim which Joyce wishes to make about the nature of morality, without arriving at the conclusion that moral discourse is fundamentally misguided in its aspirations.

The first chapter of the thesis is expository. In it, I provide an overview of modern evolutionary interpretations of human behaviour, and explain what it means to say that a particular trait has evolved. I begin with a discussion of E. O. Wilson's sociobiological research programme, which sparked much controversy when it was initiated in the1970s. Wilson's work on sociobiology laid the foundations for more recent research into evolutionary psychology. The latter is heavily influenced by its historical connections to sociobiology. Nevertheless, it is a distinct school of thought, having closer conceptual ties to modular theories of mind, and placing a greater emphasis on the adaptive discontinuity between ancestral humans' lives in the Pleistocene, and life in modern, post-industrial societies. Having set out the theoretical foundations of evolutionary interpretations of human behaviour, I show how these have been applied to the phenomenon of cultural evolution. I discuss William Durham's model of gene-culture co-evolution, as well as Peter Richerson and Robert Boyd's work on genetically adaptive cultural heuristics. This discussion shows how positing the existence of very general evolved psychological traits can generate interesting and informative models of cultural evolution.

Of course, such an approach is not without its detractors. Accordingly, following my discussion of cultural evolution, I examine the major objections which have been made against sociobiology and evolutionary psychology. These objections point to the need for an evolutionary approach which is more sensitive to the role played by environmental factors in trait development. I then argue that developmental systems theory constitutes just such an approach. A developmental perspective on trait evolution allows us to combine insights from evolutionary psychology with a

---

[6] See, for example, Blackburn (1993) and Gibbard (1990; 2003).
[7] For discussion, see Cuneo (2006).

flexible theoretical framework, which does not make problematic presuppositions about the way in which traits are selected for.

Adopting a developmental systems perspective necessitates revising one's conception of what the term "evolved" denotes, so as to include enduring, extra-genetic influences on development. In Chapter Two, I examine the issue of what constitutes an evolutionary theory of ethics, doing so from a developmental systems perspective. This issue is not a straightforward one, as many ethical theories which purport to be non-evolutionary make very similar empirical and conceptual claims to explicitly evolutionary theories. Such similarity may seem un-problematic from the perspective of evolutionary psychology. However, a developmental systems analysis shows that as the literature stands, the ways of distinguishing between evolutionary and non-evolutionary theories of ethics are deeply problematic. I argue that this distinction can be made more philosophically respectable by grounding it upon the difference between evolutionary moral philosophy and evolutionary moral psychology, rather than on speculation regarding the extent to which morality is "in our genes".

Having described how evolutionary data can be used to explain behavioural traits, and having argued for a specific way of classifying ethical theories as evolutionary, I turn in Chapter Three to Joyce's argument for an error theory. The chapter begins by highlighting some important similarities between Joyce's argument and the argument for an error theory put forward by John Mackie. With this background in place, I describe how Joyce thinks evolutionary forces have influenced our moral beliefs, and explain how this fits in with his conception of ordinary moral discourse to generate an error theory. Joyce's argument is fundamentally epistemological. Very roughly, he argues that natural selection for the possession of pro-social dispositions allows us to explain the general form of our moral beliefs, i.e., that moral properties exist, and that they generate categorically binding moral obligations. However, the evolutionary narrative which Joyce endorses does not entail the existence of moral properties. This undermines our epistemic warrant for believing in those moral properties, as our moral judgements can be effectively explained without appealing to their regulatory influence. Furthermore, Joyce argues, moral properties are not the sort of properties which can be accommodated in a naturalistic framework. This is because natural properties cannot make categorical demands on agents. We therefore have good reason not only to suspend our belief in moral properties, but to altogether deny that they exist.

Joyce's argument for an error theory therefore has two prongs: an epistemological prong, according to which evolutionary data undermines our reasons for believing in moral properties; and a metaphysical prong, according to which moral properties cannot be satisfactorily naturalised. I deal with the second prong first, by adapting an argument made by Jamie Dreier. I claim that categoricity is generated not by moral properties *themselves*, but rather by the fact that we *experience* moral reasons as categorically binding. Categoricity is a feature of agents' subjective motivational sets. I argue that Joyce is mistaken to insist that ordinary moral discourse treats moral reasons as binding on all logically possible agents, irrespective of their psychological make-up. We do not, therefore, need to give a naturalistic account of moral properties which shows them to have this feature. The account of moral categoricity which I defend is attractive because it coheres with Joyce's account of moral judgement, and so does not beg the question against him. Even accepting his account then, moral properties need not be conceived of as Joyce conceives of them. Having shown that moral properties *can* be naturalised, at least in principle, I am still left with the first

prong of Joyce's argument. Subsequent chapters are devoted to avoiding this prong. They search for a satisfactory, realist account of moral properties, according to which belief in those properties need not be undermined by evolutionary data.

Chapter Four critically discusses Philippa Foot's neo-Aristotelian approach to virtue ethics. Foot argues that possession of the virtues is a necessary part of human flourishing, at which all practically rational action aims. According to Foot's account, evaluations of moral goodness and badness operate in the same way as evaluations which describe non-moral traits, such as keenness of hearing or eyesight, as good or bad. Each such evaluation, she argues, pertains to an organism's "natural goodness";[8] i.e., its ability to engage in species-typical behaviours, characteristic of the organism in question. Foot's account suggests a way in which to meet Joyce's epistemological challenge. If moral properties are analysable in terms of their contribution to an organism's natural goodness, and if natural goodness is determined by facts about what organisms have evolved to be like, then moral properties will themselves have been generated by the evolutionary process. Thus, *contra* Joyce, an evolutionary aetiology of morality need not undermine our belief in the existence of moral properties. However, Foot's account is deeply problematic. I consider important objections which have been made against it by Joyce and by Tim Lewens, and also offer my own criticism. Foot's commitment to *eudaimonism*, and to a welfarist conception of function, make her account philosophically untenable.

Having argued that Foot's theory cannot meet the objections which have been made against it, I consider whether an alternative approach to virtue ethics fares any better. I discuss Jonathan Haidt and Craig Joseph's Darwinian account of the virtues, according to which virtuous character traits are generated by an evolved disposition to respond with positive affects towards certain features of social situations. The features towards which we positively respond are culturally malleable, but once certain affective dispositions have been established, these dispositions can be said to generate culturally specific virtues. Insofar as Haidt and Joseph's theory avoids a commitment to *eudaimonism* and to a welfarist conception of function, its philosophical foundations are firmer than those which support Foot's theory. Yet Haidt and Joseph's account is explicitly not designed to act as a normative ethical theory. Objective moral properties do not form any part of their account of the virtues. As a result, their account cannot, at least as it stands, be used to rebut Joyce's argument for an error theory.

Chapter Five discusses Jesse Prinz's brand of realist sentimentalism. Prinz's position suggests a way of reconciling the Darwinian perspective adopted by Haidt and Joseph with a metaphysically unproblematic account of moral properties. Prinz's account is self-consciously Humean. He argues that moral judgements are emotionally constituted, and draws on Haidt's work on moral psychology to support this claim. Furthermore, Prinz holds that moral properties can be cashed out in naturalistic terms. He attempts to do this by appealing to a causal notion of realism, according to which moral properties are just those properties which cause us to respond to a situation with a particular moral emotion. Although Prinz would not describe his account as an evolutionary one, his theory, if successful, would constitute a reply to Joyce's epistemological challenge. Given this, and given also its compatibility with Haidt's evolutionary perspective on moral psychology, Prinz's realist sentimentalism could be used to supplement Haidt and Joseph's account of the virtues. Doing so

[8] Foot (2001) p. 3.

would thereby generate a metaethically realist, Darwinian virtue ethic. I argue, however, that Prinz's theory faces serious difficulties. Firstly, I show that his analysis of the basic values from which our moral sentiments are derived is incompatible with his theory of emotion. Secondly, I argue that Prinz's conception of realism is insufficiently robust. His account fails to satisfy the *desiderata* for realist sentimentalism set out by Justin D'Arms and Daniel Jacobsen. That is to say, Prinz's theory cannot always be called upon to settle moral debate: the occasions on which it can do so will depend upon the extent to which disputants share the same basic values. As a result, Prinz's version of realist sentimentalism fails to satisfy one of the archetypal requirements of normative realism.

The sixth, and final, chapter argues that an evolutionary approach to care ethics issues in a more robustly realist form of sentimentalism than does Prinz's approach, and that this is in fact the best theory with which to meet Joyce's sceptical challenge. I give a brief account of the origins of care ethics in Carol Gilligan's feminist critique of developmental psychology, and outline the central tenets of care ethics, including the importance of empathy in discussions of caring. Following this, I turn my attention to moral psychology. I discuss research conducted by Hauser et al into the extent to which our moral judgements are influenced by knowledge of an agent's intentions, and argue that this research shows moral judgement to be guided by the sorts of considerations central to care ethics. I argue that the expressed moral principles which we use to rationally justify our moral judgements should be understood as culturally evolved heuristics, designed to facilitate the performance of actions which accord with our operative moral principles. These operative principles generate our moral intuitions, and our expressed moral principles are subordinate to them. Operative principles guide our judgements about the validity of the expressed principles which we formulate, and veto them in the event that they substantially depart from the content of the operative principles. These operative principles, I argue, should be interpreted specifically as care-ethical intuitions. I draw support for this argument by discussing research into the selection pressures responsible for the evolution of human intelligence. The need to predict and explain the behaviour of conspecifics was crucial to survival in early human populations, and humans evolved a sophisticated theory of mind and the ability to engage in empathetic perspective-taking in order to more successfully navigate their social environment. The ability to make inferential judgements about the motivations of particular agents' actions was a result of selection for empathy and theory of mind. I argue that such judgements form the basis of our moral intuitions, and that, so understood, these intuitions are care-ethical.

Having argued that our moral intuitions are care-ethical intuitions, and that they carry more normative weight than our expressed moral principles, I go on to show that an evolutionary care ethics constitutes a robust moral realism. I show that my approach is a form of reference-fixing moral realism, according to which the meaning of moral terms is specified by an account of their Darwinian function. I show that this approach is able to avoid the relativistic pitfalls that make other versions of sentimentalism, most notably those of Prinz and Michael Slote, objectionable. Finally, I show that the evolutionary narrative which grounds my conception of care ethics shows that evolutionary data does not give us reason to doubt the existence of moral properties. Joyce's argument for an error theory is therefore refuted.

Chapter One

Evolutionary Interpretations of Human Behaviour

1. Introduction

This expository chapter details the way in which evolutionary data have been used to analyse human behaviour. It explains how the evolutionary processes which guided human beings' phylogenetic history are thought to relate to the development of modern human societies. The following discussion therefore provides an answer to the question: "what does it mean to say that certain human beliefs or behaviours are the product of evolution?"

As will become apparent over the course of this chapter, subtly different answers have been given to this question. The first such answer to consider is that offered by E. O. Wilson. During the 1970s Wilson pioneered research into the field of sociobiology. This research sought to incorporate biological data into attempts to understand and explain human culture. Wilson's approach is the natural starting point for a discussion of evolutionary explanations of human behaviour, as it provides the conceptual foundations from which subsequent accounts have been developed.

Following discussion of Wilson's sociobiology in §2, §3 will outline its historical successor, evolutionary psychology. The main points at which evolutionary psychology agrees with and departs from sociobiology will be highlighted. A frequently discussed example of evolutionary psychology's analysis of moral judgement will then be considered in order to flesh out more fully the sort of claims made by evolutionary psychologists.

In §4, the relation of evolutionary psychology to cultural evolution will be considered. It may be thought that the traits studied by evolutionary psychologists are too basic to make their study in relation to cultural and moral norms worthwhile. Thus, if all cultures and moral systems are ultimately derived from the same cognitive adaptations, those adaptations seem unable to explain why such systems developed in uniquely different ways. The data which does explain this change, and which is of historical and philosophical interest, may therefore seem to lie outside of the scope of evolutionary psychology. By discussing the literature on gene-culture co-evolution, I argue that this is not the case. Evolutionary psychology *does* have a direct bearing on the study of cultural evolution.

With an exposition of the main tenets of evolutionary psychology and its relation to cultural evolution having been provided, §5 discusses some of the objections which have been made to that approach. Whilst none of these prove fatal to the basic approach endorsed by evolutionary psychology, they do suggest that it unduly ignores the possibility of accounting in other ways for the phylogenetic and ontogenetic development of the traits which it studies. Essentially, evolutionary psychology tends to systematically overemphasise the importance of adaptive, modular explanations. Whilst these may be essential to the study of particular traits, this must be conclusively established on a case-by-case basis, not simply assumed as the starting point of an investigation.

Finally, §6 outlines an alternative approach to evolutionary psychology, namely developmental systems theory. The latter approach is not susceptible to the criticisms which have been made with respect to evolutionary psychology. Developmental systems theory therefore provides a more compelling account of the evolution and development of traits. As will be seen, however, accepting the claims made by developmental systems theory means reconceiving the term "evolution" and the process to which it refers.

2. Sociobiology

The claim that morality, or human behaviour more generally, is in some sense the product of biological evolution is easily misunderstood. Such a claim might be thought to imply either (or both) of two theses. These are: (i) there is no such thing as genuine altruism, as everything that we do is motivated by a concern to spread copies of our genes and is therefore ultimately selfish; and (ii) all of our actions and moral judgements are genetically determined. Neither of these theses, however, forms a part of modern evolutionary interpretations of human behaviour. I postpone discussion of thesis (i) until the account of the evolution of pro-social emotions given in chapter three. Paying some critical attention to thesis (ii), however, will be a helpful way of outlining the sort of claims made by E. O. Wilson's sociobiological research programme. As Wilson's account paved the way for later evolutionary interpretations of human behaviour, it is to this that I now turn.

(i) Sociobiology and Genetic Determinism

"The question of interest is no longer whether human behavior is genetically determined; it is to what extent."[1] Remarks of this sort seem unequivocally to assert that at least some human behavioural traits are genetically hardwired: if a person possesses the gene (or genes) for a racist or violent temperament, then they will be racist or violent. There seems to be no escaping the destiny of genetic determination: genes carry the instructions for building brains, and the mind is a product of the brain and its neural networks, which are themselves "undeniably encoded in the genes".[2] All this seems to point to the conclusion that accepting the evolution of human behavioural traits also requires us to accept that such traits will inevitably develop: they are a part of human nature. This is certainly the conclusion reached by Ashley Montagu, who writes that:

> [i]nherent in the idea of sociobiology as presented by Wilson [...] is the notion of the genetic determinism of behavior of individual differences, social differences, the stratification of classes, sexual status, and racism.[3]

---

[1] Wilson (1978) p. 19.
[2] Wilson (1978) p. 55.
[3] Montagu (1980) p. 12.

Such a conclusion, if true, would of course be disheartening for anyone committed to the ideal of creating a more peaceful, egalitarian society. Such ideals, and the perceived threat which sociobiology posed to them, motivated many attempts to refute the claim that human behaviour has an evolutionary basis. For example, Steven Rose writes:

> It is easy to imagine a world better than the present [...]. Yet at the same time our imagination of, our striving for, the new world runs full tilt into the claims of "hard-nosed realism". What is, is what must be. It's only human nature. Offered a vision of Utopia, the realist defenders of the status quo substitute sociobiology. [...] So it is important to look at the method and reasoning employed by this law-giving subject which claims to tell us who we are and how we must live.[4]

However, when interpreted as a purely politico-moral objection, Montagu's assertion that sociobiology leads to genetic determinism does nothing to call the legitimacy of such research into question. If sociobiology were to make such claims, they could not be shown to be false simply because they were politically unwelcome.

Yet there is another way of interpreting Montagu's objection, and on this alternative approach it *does* raise a problem for sociobiology. If sociobiology were committed to the claim that differences between people's moral beliefs were solely explicable by appeal to genetic differences between those individuals, this would cast serious doubt on its veracity. This is because such a claim would be far too strong to be empirically plausible. It would entail that an individual's cultural background, education, religious beliefs, and upbringing played no role in the formation of her moral character. Were it committed to this position, Wilson's sociobiology would surely be far too radical a thesis to take seriously.

In fact, as noted above, sociobiology does *not* claim that certain types of behaviour or particular character traits are genetically determined. When Wilson writes that the question is the *extent to which* human behaviour is genetically determined, he simply means that human genes will *influence* the development of behavioural traits. For Wilson, the development of any such trait is the outcome of an interaction between evolved cognitive predispositions and a person's developmental environment. "Each person is molded by an interaction of his environment, especially his cultural environment, with the genes that affect social behavior."[5] More specifically, Wilson claims that the human mind should be understood as:

> an autonomous decision making instrument, an alert scanner of the environment that approaches certain kinds of choices and not others in the first place, then innately leans toward one option as opposed to others and urges the body into action according to a flexible schedule that shifts automatically and gradually from infancy into old age.[6]

The innate preferences to which Wilson alludes in the above quotation are elsewhere said to develop through the operation of "epigenetic rules". These are:

---

[4] Rose (1980) p. 160.
[5] Wilson (1978) p. 18.
[6] Wilson (1978) p. 67.

> genetically determined procedures that direct the assembly of the mind, including the screening of stimuli by peripheral sensory filters [...] and the deeper processes of directed cognition.[7]

This rather abstract characterisation can be made somewhat clearer by considering specific behavioural traits such as aggression. Thus Wilson argues that:

> it is true that aggressive behavior, especially in its more dangerous forms of military action and criminal assault, is learned. But the learning is prepared [as described above] [...]; we are strongly disposed to slide into deep, irrational hostility under certain definable conditions.[8]

The above quotations certain show that Wilson's approach raises significant questions about the potential malleability of human psychology. Nevertheless, it is not committed to empirically implausible claims about the genetic *determination* of psychological traits. Rather, it sees such traits as emerging from interactions between an individual's environment, and her innate dispositions to react to that environment in specific ways. By changing the environmental stimuli with which an individual interacts, therefore, it is at least theoretically possible to change the psychological and behavioural traits which she develops. As a result, Wilson felt able to express the hope that his position "should satisfy both camps in the venerable nature-nurture controversy".[9] As has already been seen, this statement was somewhat over-optimistic.

A new question now arises, however: how plausible is Wilson's claim that we possess various innate dispositions that incline us towards developing specific psychological traits?

(ii) Learning Biases

There is in fact a compelling amount of both theoretical and empirical evidence to support Wilson's claim that there are innate learning biases.

On the theoretical front, Lumsden and Wilson (1981) argue that a population consisting of individuals without such learning biases would not exist in an evolutionarily stable state. Roughly, an evolutionarily stable state is one in which a population of organisms have adopted some behavioural strategy, such that other organisms in the population who do *not* behave according to the same strategy suffer a comparable loss in genetic fitness and are selected against. This definition will be clarified by considering Lumsden and Wilson's argument that a population without any innate learning biases is evolutionarily unstable.

This argument is developed with reference to the concept of "culturgens". These are simply the "array of transmissible behaviors, mentifacts [i.e. concepts], and artifacts"[10] which an organism

---

[7] Lumsden and Wilson (1981) p. 7.
[8] Wilson (1978) p. 106.
[9] Wilson (1978) p. 105.
[10] Lumsden and Wilson (1981) p. 7.

can acquire. Lumsden and Wilson argue that a population with no innate disposition to respond favourably to adaptive culturgens would be vulnerable to displacement by a population that *did* possess such biases. Because the existence of innate biases is determined by the operation of epigenetic rules, and because the epigenetic rules are genetically determined, there will be selection pressure in favour of organisms with innate learning biases. Lumsden and Wilson summarise their position thus:

> consider a population of *tabula rasa* organisms [...] [whose] developmental field is flat. The population is exposed to an environment that contains both adaptive and deleterious culturgens, but it is unable to distinguish them. [...] Over a period of generations, the population is unstable against invasion by genetic mutants that program epigenetic rules biasing individuals towards assimilation of relatively adaptive sets [i.e. of culturgens]. The epigenetic rules will then tend to channel cognitive development toward certain culturgens as opposed to others.[11]

Lumsden and Wilson's argument makes good evolutionary sense. It is also supported by empirical observations. For example, Marc Hauser reports observations which suggest that rhesus monkeys have innate learning biases pertaining to the recognition of fearful stimuli.

> If a group of rhesus monkeys with no snake experience watches an experienced group express fear toward [a] snake, the observers will readily absorb this fear, responding with alarm the next time they confront the snake. In contrast, if a naïve group of rhesus watches other rhesus show fear toward a bed of flowers, the fear doesn't spread; the next time they confront a bed of flowers, there is no fear at all. It would take a lot more to convince the primate mind that flowers count [i.e. as fearful].[12]

Humans also exemplify the sort of learning biases found in rhesus monkeys. Matt Ridley observes that, in addition to snakes,

> [p]eople are also commonly afraid of spiders, the dark, heights, deep water, small spaces and thunder. All these were a threat to Stone Age people, whereas the much greater threats of modern life – cars, skis, guns, electric sockets – simply do not induce such phobias.[13]

Ridley concludes that "the human brain is pre-wired to learn fears that were of relevance in the Stone Age".[14]

Of course, this is not to suggest that such fears will *inevitably* develop, or that a fear of cars, guns or electric sockets simply *cannot* be instilled in someone. There are plenty of phobias which cannot be understood as innate responses to persistent threats found in humans' environment of evolutionary adaptation: triskaidekaphobia, i.e. fear of the number thirteen, is just one example among many possible others. Yet it is not being claimed that there is anything inevitable about developing certain phobias. All that is being argued here is that people have a predisposition which

---

[11] Lumsden and Wilson (1981) p. 13.
[12] Hauser (2006) p. 355.
[13] Ridley, Matt (2003) p. 194.
[14] Ridley, Matt (2003) p. 194.

makes them more likely to develop some phobic responses and not others. The prevalence of arachnophobia, and the comparative scarcity of triskaidekaphobia, makes this conclusion hard to resist.

Lumsden and Wilson's theory predicts the existence of learning biases based on the evolutionary instability of a population which lacks them. Empirical research into fear responses supports their prediction. Furthermore, once it has been conceded that there are learning biases which influence our fear responses, there is no reason, at least in principle, to deny that similar biases may also influence other areas of human cognition. According to Ruse and Wilson (1986), our moral judgements are also guided by epigenetically regulated biases.

(iii) Epigenetic Rules and Moral Judgements

The strategy which Ruse and Wilson use to explain the existence of innate biases in our moral judgements recapitulates the approach used in the discussion of aggression and fear responses, though in this instance greater attention is paid to the cultural reinforcement of moral judgements.

> The empirical heart of [the] discussion is that we think morally because we are subject to appropriate epigenetic rules. These predispose us to think that certain courses of action are right and certain courses of action are wrong. [...] The full sequence in the origin of morality is therefore the following: ensembles of genes have evolved through mutation selection within an intensely social existence over tens of thousands of years; they prescribe epigenetic rules of mental development peculiar to the human species; under the influence of the rules certain choices are made from among those conceivable and available to the culture; and finally the choices are narrowed and hardened through contractual agreements and sanctification.[15]

What sort of moral judgements will a process such as this one make developmentally probable? In answer to this question, Ruse and Wilson give the example of an innate disposition to avoid sibling incest. They claim that because children born as a result of sibling incest have a reduced genetic fitness, there has been selection in favour of an innate disposition to avoid such incest. However, "this biological cause and effect is not perceived in most societies, especially those with little or no scientific knowledge of heredity".[16] Thus, whilst detrimental fitness effects are the ultimate explanation for the existence of an incest avoidance bias, an awareness of these effects cannot be what regulates that bias. Rather, such regulation is achieved by a blanket avoidance of sexual relations between children who, "between birth and approximately six years [of age] [...] are exposed to each other under conditions of close proximity".[17]

---

[15] Ruse and Wilson (1986) pp. 180 – 181.
[16] Ruse and Wilson (1986) p. 184.
[17] Ruse and Wilson (1986) p. 184.

Why does the incest avoidance instinct develop in this way? According to Debra Lieberman, it does so because, in the environment of evolutionary adaptation, this was the most reliable way of tracking degrees of familial relatedness. As she explains:

> genes were not available for direct comparison in ancestral environments. Therefore, kin detection would have had to rely on *cues* that highly correlated with patterns of underlying genetic relatedness.[18]

Childhood co-residence is likely to be just such a cue, she argues.

Of course, co-residence is not an infallible guide to kin detection. In some circumstances, for example, non-kin may be raised in co-residence with one another. Under these conditions, a kin detection mechanism which tracked co-residence would misidentify an unrelated individual as a close family member. An aversion towards having sexual relations with that person would then be likely to develop, despite their explicitly recognised status as non-kin. In fact, this phenomenon seems to have been observed in at least two studies. These studies focussed on children who were raised together, either in Israeli kibbutzim, or in Taiwanese minor marriages. In the latter, unrelated young children are raised together with the explicit intention that they should later marry one another. As Lieberman reports:

> early childhood association led to a decreased probability of marriage in Israeli kibbutzim [...] and reduced rates of fertility combined with increased rates of divorce and extramarital affairs in Taiwanese minor marriages.[19]

According to Wilson's sociobiological approach, then, human behaviour is regulated by genetically determined, epigenetic rules. These rules were produced by natural selection, which operated so as to create agents disposed to act in ways that promoted their genetic fitness. There is both theoretical and empirical evidence to support this claim. Furthermore, moral judgements are no exception to this rule. Incest avoidance is just one example of an innate moral judgement of the sort predicted by sociobiology. In principle, there may be many other innate moral biases. For example, Ruse and Wilson (1986) suggest that there may be an epigenetic rule disposing us to be highly concerned with issues of justice and fairness in social exchange,[20] although they do not elaborate on this idea. The next section discusses how this suggestion has been developed from the perspective of evolutionary psychology.

3. Evolutionary Psychology

(i) The Difference Between Sociobiology and Evolutionary Psychology

[18] Lieberman (2008) p. 179.
[19] Lieberman (2008) p. 176.
[20] Ruse and Wilson (1986) p. 185.

Sociobiology provided the basic theoretical framework for modern, evolutionary interpretations of human behaviour. Such interpretations are now typically characterised as belonging to the field of evolutionary psychology (hereafter EP). As seen in the previous section, sociobiology's claims were often taken to have deeply unwelcome political implications. As Wilson describes it, his work sparked a "controversy stemming from [a] mix of misunderstanding, suspicion, and resentment".[21] In the light of this controversy, EP has sometimes been described as a politically sanitised rebranding of sociobiology. Indeed, this is a view which Wilson himself has endorsed:

> in the interest of simplicity, clarity, and – on occasion – intellectual courage in the face of ideological hostility, evolutionary psychology is best regarded as identical to human sociobiology.[22]

It is certainly true that there are important historical and conceptual links between sociobiology and EP. This is because EP developed out of sociobiology, and elaborated, rather than rejected, much of its theoretical content. Thus both take natural selection to have equipped the human mind with innate psychological biases, and both take these biases to be adaptations to humans' ancestral environment.

However, despite Wilson's remarks to the contrary, sociobiology and EP are in fact conceptually distinct schools of thought. The most important difference between the two lies in EP's development of a modular account of human cognition. Sociobiology, it will be recalled, limited its evolutionary analysis to predictions about the sort of innate biases likely to be present in human psychology. EP extends this analysis to include discussion of *the way in which* those biases are likely to have been selected for. This seemingly subtle shift in explanatory focus generates a fundamentally different conception of the human mind and its operation.

For sociobiology, dispositions towards certain sorts of behaviour will have been targeted by selection if such behaviour was evolutionarily adaptive. Accordingly, this approach suggests that there will be general and consistent selection pressure in favour of developing *any* epigenetic rule which makes the adoption of an adaptive culturgen more likely.

The evolutionary interpretations of behaviour provided by EP offer a much more detailed account of how selection brings about innate psychological dispositions. For evolutionary psychologists, the human mind is the product of distinct, targeted selection pressures. Any innate dispositions must be understood, *not* as the result of selection for adaptive behaviours *per se*, but of selection for adaptive solutions to *specific challenges* posed by the environment of evolutionary adaptation. Such challenges will have been faced at different points in time throughout mankind's evolutionary history, and their various solutions will have required the development of different types of behavioural response. As a result, evolutionary psychologists hold that evolution will have given rise to predominantly modular, domain-specific forms of cognition. As John Tooby and Leda Cosmides explain:

---

[21] Wilson (1978) p. xvi.
[22] Wilson (1998) p. 165.

> [d]ifferent adaptive problems are often incommensurate. They cannot, in principle, be solved by the same [cognitive] mechanism. To take a simple example, the factors that make a food nutritious are different from those that make a human a good mate or a savannah a good habitat.[23]

Biological evolution takes place at a very gradual pace. For this reason, evolutionary psychologists typically see the human mind as still being adapted primarily to the conditions in which it evolved.

> What we think of as all of human history – from, say, the rise of the Shang, Minoan, Egyptian, Indian, and Sumerian civilizations – [...] are all the novel products of the last few thousand years. In contrast to this, our ancestors spent the last two million years as Pleistocene hunter-gatherers, and, of course, several hundred million years before that as one kind of forager or another. These relative spans are important because they establish which set of environments and conditions defined the adaptive problems the mind was shaped to cope with: Pleistocene conditions, rather than modern conditions.[24]

Two important predictions follow from EP's account of the human mind. The first is that any psychological traits which are the products of evolved mental modules will be highly likely to be expressed. In technical terms, they will be "developmentally canalised". The concept of developmental canalisation stems from the work of Conrad H. Waddington. Waddington likened the development of a trait to the trajectory of a ball rolling down a series of slopes, each of whose sides possess a different gradient. The ball's trajectory can sometimes be altered as a result of environmental influences, but this will not always be the case. Some of the developmental channels have such a steep gradient that no amount of environmental pressure can perturb the development of the trait which they produce. When the development of a trait is "adjusted so as to bring about one definite end-result regardless of minor variations in conditions during the course of [development]"[25] the trait is said to be canalised. Wilson provides an example of this process which helps to clarify Waddington's description:

> An individual can end up either right- or left-handed. If he starts with the genes or other early physiological influences that predispose him to the left hand, that branch of the developmental channel can be viewed as cutting the more deeply. If no social pressure is exerted, the ball will in most cases roll on down into the channel for left-handedness. But if parents train the child to use the right hand, the ball can be nudged into the shallower channel for right-handedness.[26]

Because EP sees the evolved mind as modularly constructed, evolved psychological traits will be isolated from domain-general learning. The outputs of those modules will therefore typically be uninfluenced by other developmental processes. EP therefore differs from Wilson's sociobiology, in that it significantly downplays the relevance of exposure to environmental stimuli in the acquisition of evolved traits.

---

[23] Tooby and Cosmides (1992) p. 111. Citation removed.
[24] Cosmides, Tooby and Barkow (1992) p. 5.
[25] Waddington (1942) p. 563.
[26] Wilson (1978) p. 61.

The second prediction made by EP is that some evolved psychological traits may prove to be genetically maladaptive in modern societies. It is important to note here that in this context "modern" does not only refer to post-industrial societies. It refers to *any* society which is no longer oriented around hunting and gathering. Still, the further removed a society's cultural practices are from those found in hunter-gatherer communities, the more likely it is that at least some of those practices will prove to be maladaptive.

A helpful example of an evolved psychological trait which is maladaptive in post-industrial societies is that of a preference for foods which were sugary, salty, or fatty. In the environment of evolutionary adaptation, such foods were a valuable source of nutriment. They were also hard to come by. Selection would therefore have favoured those organisms which developed a marked preference for these foods as opposed to less nutritious alternatives. In the modern world, however, sugary, salty, and fatty foods are readily available. Our evolved preference for a diet high in these substances can now lead to their overconsumption, to the detriment of an individual's health. This psychological trait has therefore become maladaptive due to the environmental disparity between foraging conditions in the Pleistocene and the abundance of food in the modern world. Furthermore, despite our awareness of the risks to health posed by their overconsumption, we still find foods containing a relatively high proportion of these substances appealing. According to EP, this is because our innate preference for these foods is the product of a domain specific module, and is therefore developmentally canalised: because it was not learned in the first instance, it cannot now be *un*learned.

There is, then, an important difference in the details of the claims made by EP and those made by sociobiology. Unlike sociobiology, EP endorses an explicitly modular theory of mind. Those modules are seen as adaptive solutions to environmental challenges, posed by conditions faced by our ancestors during the Pleistocene. Lastly, the output of those mental modules is taken to be developmentally canalised. With this brief account in place, it is now time to consider how EP approaches the evolution of morality.

(ii) The Evolutionary Psychology of Morality

It was noted in §2.iii that Ruse and Wilson (1986) predicted the epigenetic regulation of attention to fairness in social exchange. In evolutionary psychological terms, this becomes the prediction that "the recurrent structure of [social exchange's] characteristic problem type, as encountered under Pleistocene conditions,"[27] will have created sufficient selection pressure for the evolution of a cognitive solution to that problem. This "characteristic problem type" is to be understood as the possibility of missing out on potential benefits as a result of engaging in an inequitable social exchange.

Because of the unique adaptive challenges posed by this scenario, a cognitive solution specific to it will have been the target of selection. Cosmides and Tooby argue that, owing to this specifically targeted selection pressure:

[27] Cosmides and Tooby (1992) p. 166.

> one expects cognitive adaptations specialized for reasoning about social exchange to have some design features that are particular and appropriate to social exchange, but that are not activated by or applied to other content domains.[28]

They designed a test to check for the existence of such a cognitive module. This test was based on the Wason selection task, and measured subjects' ability to test for the violation of a conditional rule. Performance on such tasks is typically poor, and is only slightly improved by increasing subjects' familiarity with the conditional rules being tested. Even when subjects are highly familiar with the conditional rule in question, the logically correct response is still given by fewer than half of those tested.[29] Thus "[h]umans do not appear to be naturally equipped to seek out violations of descriptive or causal [conditional] rules".[30]

Cosmides' and Tooby's test asked subjects to check for the violation of a conditional rule with an "If P then Q" logical structure. This conditional was expressed both abstractly, and in the form of a social contract.[31]

In its abstract formulation, the subjects were asked to test the rule that documents with a "D" rating were marked with code "3". Subjects were then presented with four cards marked "D", "F", "3", and "7" respectively. They were instructed to identify the two cards which needed to be turned over in order to check for a violation of the rule (in this case, "If D then 3").

In its social contract formulation, the subjects were asked to test the rule that beer drinkers must be over the age of twenty. Subjects were presented with four cards, which were marked "drinking beer", "drinking soda", "25 years old", "16 years old" respectively. Again, subjects were asked to identify the two cards which would need to be turned over to check for a violation of the rule (in this case, "If drinking beer, then over twenty").

The correct solution to each test is the same. Subjects ought to turn over cards marked "D" and "drinking beer", and cards marked "7" and "16 years old". That is, they check for conditions in which P applies, but Q is violated.

Cosmides and Tooby found that subjects' ability to correctly test for the violation of conditional rules significantly increased when these rules were given in the form of a social contract. To test whether this result was due to the greater familiarity of the social contract form of the rule, the same test was repeated with social rules that were unfamiliar to test subjects. It was found that "the performance level for unfamiliar social contracts is just as high as it usually is for familiar social contracts such as the drinking age problem".[32] Based on these results, Cosmides and Tooby conclude that "human reasoning changes dramatically depending on the subject matter one is reasoning about".[33] More specifically, they claim that:

---

[28] Cosmides and Tooby (1992) p. 166.
[29] See Cosmides and Tooby (1992) p. 181.
[30] Cosmides and Tooby (1992) p. 183.
[31] The following summary has been adapted from Cosmides and Tooby (1992) p. 180.
[32] Cosmides and Tooby (1992) p. 186.
[33] Cosmides and Tooby (1992) p. 183.

> [t]he results showed that we do not have a general-purpose ability to detect violations of conditional rules. But human reasoning is well designed for detecting violations of conditional rules when these can be interpreted as cheating on a social contract.[34]

This conclusion, and the data which supports it, provides strong empirical support for EP's claim that humans possess a mental module dedicated to assessing the fairness of social exchange.

The social exchange module postulated by Cosmides and Tooby operates at a very basic level. It neither generates specific moral beliefs, nor determines how specific beliefs pertaining to social exchange will develop within a culture. Its account of social exchange practices is therefore much less detailed than the epigenetic account of incest avoidance discussed in §2.iii. This generality can be considered a virtue, insofar as it allows EP to offer a unified (or rather, unifying) analysis of a broad range of cultural phenomena.

> From the child who gets her dessert if her plate is cleaned, to the devout Christian who views the Old and New Testaments as covenants arrived at between humans and the supernatural, to the ubiquitous exchange of women between descent groups among tribal peoples [...] all of these phenomena require, from the participants, the recognition and comprehension of a complex set of implicit assumptions that apply to social contract situations. Our social exchange psychology supplies a set of inference procedures that fill in all these necessary steps, mapping the elements in each exchange situation to their representational equivalents within the social contract algorithms, specifying who in the situation counts as an agent in the exchange, which items are costs and benefits to whom, who is entitled to what, under what conditions the contract is fulfilled or broken, and so on.[35]

This level of generality comes at a cost, however. Philosophers interested in the plausibility of evolutionary interpretations of human behaviour may be tempted to conclude that such explanations are in fact mere pseudo-explanations. If the same analysis is offered by way of an explanation for phenomena as diverse as those mentioned in the above quotation, to what extent is our understanding of any one of those practices really improved?

These are legitimate concerns. Yet in fact EP can provide the basis for a more detailed, and thus more intellectually satisfying, explanation of cultural phenomena. To see how this can be done, it is necessary not only to consider the sort of evolved psychological traits which we as a species possess, but also how these effect the development of culture. This will be the task of the next section. Following that discussion, we will consider some of the objections which have been made against EP. This will pave the way for the subsequent endorsement of a less problematic alternative to EP, namely, developmental systems theory.

4. Co-evolutionary Theories of Culture

[34] Cosmides and Tooby (1992) p. 205.

[35] Cosmides and Tooby (1992) p. 207.

Research into evolutionary anthropology has provided an important supplement to the sort of claims made by sociobiology and EP. The accounts discussed in this section are commensurate with the evolutionary analyses of human behaviour given by these theories, yet seek to more fully explain how evolved psychological biases (whether these are domain-specific or otherwise) influence cultural change. As William Durham explains:

> It appeared to me not so much that sociobiology was wrong [...] as that sociobiology was intrinsically incomplete as a theory for explaining human behavioral diversity.[36]

The same sentiment is also expressed by Peter Richerson and Robert Boyd, when they write that evolutionary psychologists:

> are surely right in stating that every form of learning, including social learning, requires an information-rich innate psychology, and that much of the adaptive complexity we see in cultures around the world stems from this information. However, ignoring transmitted culture completely is a big mistake. [...] Cumulative cultural adaptation cannot be based directly, or in detail, on innate, genetically coded information.[37]

Attending to the way in which these authors supplement sociobiological and evolutionary psychological analyses will therefore dispel the concern, raised at the end of the previous section, that evolutionary explanations of human behaviour are mere pseudo-explanations.

(i) Memes and Cultural Selection

Durham's influential account of the relationship between genetic and cultural evolution describes the process of cultural change in terms of "the transmission of socially meaningful information".[38] Following Dawkins (1982), Durham characterises this information in terms of units specifiable as "memes". According to Durham, then, "whenever culture changes, *some* ideational unit [i.e. a meme] is adopted and one or more homologous alternatives are not".[39] "Meme" must therefore be understood very broadly, including phenomena as diverse as words, concepts, or practices. Durham divides the range of alternative forms for each such meme into two categories: "holomemes" and "allomemes". The former comprise "the entire cultural repertory of variation for a given meme, including any latent or unexpressed forms".[40] The latter constitute "the subset of holomemes that are actually used as guides to behavior by at least some members of a population in at least some circumstances".[41] Durham makes this rather vague description more vivid by providing some examples. Thus the term "allomeme" can pick out:

---

[36] Durham (1991) p. 2.
[37] Richerson and Boyd (2005) p. 45.
[38] Durham (1991) p. 188.
[39] Durham (1991) p. 189.
[40] Durham (1991) p. 189.
[41] Durham (1991) p. 189.

> differing techniques or strategies for procuring subsistence resources; alternative schools or sects of religious thought co-existent within a population; differing conceptions about the length of a postpartum sex taboo; or variable definitions of a word or label [...].[42]

"Holomeme", by contrast, picks out not only the variable definitions of a word actually found within a population, for example, but also extends to cover all logically possible alternative definitions of that word, including those which have never been adopted.

With this terminology in place, Durham goes on to provide an account of the nature of the relationship between genetic and cultural evolution. He does this by way of a discussion of what he terms "cultural selection". Cultural selection is to be understood as "the differential social transmission of cultural variants through human decision making, or simply as 'preservation by preference'".[43] This process of cultural selection is mediated by the influence exerted upon human decision making by our "primary" and "secondary" values.

Secondary values are explicitly cultural. They are socially transmitted phenomena, i.e. memes, "arising from collective experience and social history".[44] Importantly for the focus of the present discussion, Durham characterises moral principles as secondary values. These also include phenomena such as folk sayings, local customs, and conventions pertaining to etiquette: anything which is used by individuals as a practical guide to daily life. Secondary values therefore give cultural selection an important, self-regulatory aspect: some secondary values will be selected specifically because they cohere with other, previously selected secondary values.

However, secondary values are not the only driving force behind cultural selection: primary values also have an important part to play in this process. Primary values are composed of "feedback from the senses, from the internal reward system of the brain, and from organically evolved cognitive processes".[45] They therefore lack the ideational content of the secondary values, operating at a far more intuitive level. Thus, primary values are not memes, and are not culturally transmitted. Rather, they "develop within each individual out of the interaction between nervous system and environment".[46] Clearly then, Durham's primary values are precisely the sort of phenomena studied by EP and sociobiology. How, then, do primary values effect the process of cultural change?

Because primary values are not memes, any influence they exert on decision making will be pre-reflective. They simply "classify some sensations as pleasant or satisfying and others as unpleasant or frustrating".[47] Primary values are therefore to be understood as an evolved means of guiding an organism's actions towards adaptive behaviours. Despite culture's capacity for self-regulation, however, the influence of the primary values upon secondary value selection is ubiquitous. Indeed, Durham argues that secondary values must in fact be understood as a *product* of the primary value system. This is because, in the early stages of secondary value selection, a

---

[42] Durham (1991) pp. 189 – 190.
[43] Durham (1991) p. 198.
[44] Durham (1991) p. 201.
[45] Durham (1991) p. 200.
[46] Durham (1991) pp. 200 – 201.
[47] Durham (1991) p. 178.

secondary value will have been evaluated as good just in case it cohered with feedback from the primary value system. Thus, secondary values, at least in their earliest forms, can be understood as:

> true "surrogate values", helping the system to make decisions more efficiently or more reliably, but still in general agreement with what was effectively the decision criterion of genetic selection.[48]

As secondary values proliferated, they will "have been increasingly assessed by other secondary values".[49] As a result of this process, they will have developed so as to become relatively autonomous of the primary values. Nevertheless, primary value feedback will remain an on-going process. Although this feedback may not significantly influence selection between many different allomemes, it will "automatically tend to favor the preservation of any variant more consistent with the effective criterion of genetic selection".[50]

It is now possible to see how the causal influence of the primary values can explain why a set of allomemes is not co-extensive with that of the corresponding holomemes. This is because some of the logically possible holomemes are so opposed to the primary values that their unforced adoption is made massively unlikely. The influence of the primary values allows Durham to conclude that unforced cultural change will tend to be genetically adaptive. Of course this is by no means a certainty, as some changes may be adaptive in the short-term, but lead to unforeseen maladaptive consequences in the long-term. Nevertheless, genetically maladaptive cultural change is a contingency which "the human decision system has been, in effect, designed by selection to resist".[51]

The continuous affective feedback provided by the primary values can therefore be said to influence, though not to determine, the outcome of cultural selection. Still, this influence at present appears rather nebulous and imprecise. How, if at all, can something as basic as a primary value be said to have a discernible influence on selection between what Durham terms "higher-order"[52] secondary values, i.e. those which produce no obvious difference in genetic fitness?

The operation of this influence can be made more explicit by considering the account of cultural evolution given by Richerson and Boyd. Doing so will allow us to see how primary values do not simply constitute simple behavioural responses, such as finding snakes fearful or incest unappealing. Rather, they also include particular cognitive heuristics which bias the course of cultural selection. These heuristics then generate striking amounts of cultural diversity as a result of operating in different environmental contexts. An appreciation of this process will significantly disarm the worry, raised at the end of the previous section, that EP cannot provide a genuinely informative explanation of human behaviour.

(ii) Heuristics for Cultural Selection

---

[48] Durham (1991) p. 208.
[49] Durham (1991) p. 208.
[50] Durham (1991) p. 209.
[51] Durham (1991) p. 457.
[52] Durham (1991) p. 208.

Richerson and Boyd argue that cultural change (in Durham's terms, "cultural selection") is guided by general purpose rules of thumb, used by individuals to make complex behavioural decisions quickly and easily. Drawing on the work of Todd and Gigerenzer (2000),[53] they term these rules of thumb "heuristics". The use of these heuristics is likely to be widespread, as they greatly facilitate choosing between mutually exclusive allomemes. Human actions, and the consequences to which they give rise, are often highly complex. Furthermore, human beings' cognitive powers are limited. When faced with adopting one of two competing allomemes, trying to establish *de novo* which of these is likely to be most rewarding can be very challenging. Yet such choices are typically not made between only two rival alternatives, but between many. Heuristics make this selection process much easier. They are "imitation strategies [...] for guessing the right thing to do in a complex and variable environment".[54]

It is a feature of such heuristics that acting on them typically brings about results which are at least as good as, and sometimes better than, the results which would have been brought about by acting after prolonged deliberation. For example, a highly reliable heuristic for correctly identifying which of two cities is larger than the other is "opt for the city whose name you recognise the most".[55] This is a much quicker and simpler approach than possible alternatives, such as consulting a map, or finding a colleague who has been to both cities and asking them for advice.

Heuristics affect cultural evolution by biasing the transmission of social information. That is, they incline their users to "preferentially adopt some cultural variants rather than others".[56] As established in §4.i, primary values play a fundamental role in the evaluation of rival allomemes. Richerson and Boyd's discussion of cognitive heuristics in their account of cultural evolution allows us to identify one of the ways in which this influence is mediated. An agent's primary values include certain cultural selection heuristics. These, in turn, influence the likelihood that certain allomemes will be socially transmitted rather than others; i.e. they create learning biases.[57]

Richerson and Boyd go on to identify two plausible cultural selection heuristics. These are "imitate the common type"[58] and "imitate the successful".[59] A brief discussion of each will highlight how selection according to the same heuristics can nevertheless produce significant cultural divergence.

The "imitate the common type" heuristic is a "conformist bias",[60] in that it preferentially selects the most widespread cultural variant for imitation. It does this regardless of the content of that variant. Why might a bias of this sort be an expedient heuristic? Richerson and Boyd answer this

---

[53] A discussion of this research takes place in the final chapter of this thesis.
[54] Richerson and Boyd (2005) p. 118.
[55] See Richerson and Boyd (2005) pp. 119 – 120 for discussion.
[56] Richerson and Boyd (2005) p. 68.
[57] I have retained the term "allomeme" and its cognates in this section for the sake of terminological consistency. It should be noted, however, that whilst Richerson and Boyd follow Durham's terminology in speaking of primary and secondary values, they do not use the term "meme", preferring instead to talk of "cultural variants". See Richerson and Boyd (2005) p. 63 for discussion.
[58] Richerson and Boyd (2005) p. 120.
[59] Richerson and Boyd (2005) p. 124.
[60] Richerson and Boyd (2005) p. 121

question by appealing to a thought experiment. Their discussion focusses on a hypothetical population of early humans. This population, they argue, will need to develop adaptive behavioural responses to a series of environmentally specific challenges. These challenges will pertain to issues such as the optimal size of the community, how best to distribute food supplies, etc. Richerson and Boyd argue that:

> once a peripheral [...] population is isolated enough that adaptive processes cause the best variants to be most common, those who imitate the most common variant are less likely to acquire inappropriate beliefs than those who imitate at random. If this conformist tendency is genetically or culturally heritable, it will be favoured by natural selection.[61]

Richerson and Boyd have run mathematical modelling experiments to test whether selection for a conformist bias occurs, and have found that:

> selection favours a strong conformist tendency even when there is only a modest reliance on social learning. Thus, the psychology of social learning should plausibly be arranged so that people have a strong tendency to adopt the views of the majority of those around them.[62]

A conformity bias of this sort has in fact been observed in psychological experiments. For example, a subject asked to observe a light on a screen and to estimate how far the light moves can be induced to significantly alter their original estimate. This occurs when other test subjects (who are in fact stooges) give consistent estimates different to that of the subject.[63]

A second heuristic whose existence Richerson and Boyd predict is "imitate the successful". As its name suggests, this is a bias which disposes individuals to imitate successful members of their community. Such imitation is not always a straightforward matter of adopting the success-making traits of a prestigious individual, however. These traits may not always be easy to discern, and in some cases may even be unrecognised as such by the prestigious individual herself. Imitation of successful members of the community may therefore manifest itself in the indiscriminate adoption of all their imitable behavioural traits.

> Determining *who* is a success is much easier than determining *how* to be a success. By imitating the successful, you have a chance of acquiring the behaviors that cause success, even if you do not know anything about which characteristics of the successful are responsible for their success. [...] Even when the exact behaviors that contribute most to fitness are very hard to evaluate, there may be easily observable traits that are correlated with fitness, such as wealth, fame, and good health. If so, you can try to imitate everything that wealthy people do in an effort to acquire the traits that make them wealthy, but without actually trying to determine exactly how wealth is produced.[64]

---

[61] Richerson and Boyd (2005) p. 121.
[62] Richerson and Boyd (2005) p. 122.
[63] See Richerson and Boyd (2005) pp. 122 – 123 for discussion.
[64] Richerson and Boyd (2005) p. 124.

To support their claim that some such bias is likely to be at work in human psychology, Richerson and Boyd turn once again to the findings of psychological research. They note that test subjects display a marked tendency to adopt the opinions of a prestigious individual, even when those opinions are in no way related to the subject matter with which that individual's prestige is associated.

> In one study, for example, subjects were asked their opinions on "student activism" in one of three scenarios: after hearing the opinion of somebody identified as an expert on the topic, after hearing the opinion of an expert on the Ming dynasty, and after a control condition in which they didn't hear anybody's opinion. Subjects tended to voice opinions similar to either of the two experts, and were equally likely to adopt the opinions of experts on activism and the Ming dynasty.[65]

Although "imitate the common type" and "imitate the successful" are the only learning biases which Richerson and Boyd discuss in any detail, it is likely that they will interact with numerous other biases of a similar nature. There are many questions which can be asked of the account which Richerson and Boyd sketch. What is the relative influence of each of these learning biases? That is, will "imitate the successful" always trump "imitate the common type", or vice versa? Is this influence constant, or subject to environmentally regulated variation? For example, it may typically be more adaptive to imitate the common type despite the presence of relatively more successful outliers, as this avoids the risk of departing from tried-and-tested methods. However, when the relative success of these outliers rises above a certain threshold, they may come to have a disproportionate influence (relative to the rest of their community) on the distribution of material goods. At this point, switching heuristics might become the most adaptive strategy. These are not questions which I will attempt to address here. However, they demonstrate that the interaction between cultural heuristics, even when the number of these heuristics is small, can quickly become highly complex.

Learning biases of this sort, and the interactions between them, therefore create the potential for massive amounts of cumulative cultural variation. As Richerson and Boyd observe, "[s]mall, dull effects at the individual level are the stuff of powerful forces of evolution at the level of populations".[66]

It is therefore possible to allay the concern that EP is unable to provide a genuinely informative account of human cultural diversity. Once it is coupled with an analysis of the interaction between our evolved psychological dispositions and the differential selection of cultural variants, EP provides the conceptual foundation for a powerful explanatory framework. Evolutionary interpretations of human behaviour are therefore not to be dismissed as mere pseudo-explanations.

This is by no means the only philosophical objection which has been raised in response to the claims made by EP, however. Having set out the basic principles of EP, and how these purport to explain human behaviour, it is now time to consider some of those objections.

---

[65] Richerson and Boyd (2005) p. 125.
[66] Richerson and Boyd (2005) p. 123.

5. Objections to Evolutionary Psychology

This section will explore some of the objections which can be made against EP's explanatory framework. I will argue that none of these represent insurmountable challenges to EP's general approach. They do, however, highlight a pertinent criticism of EP. This is that its aetiological claims often downplay the relevance of the developmental processes which give rise to traits. Such attention to development is especially desirable when analysing traits as complex as psychological and behavioural dispositions. In §6 I will argue that the best way to guarantee providing an appropriately detailed aetiology is to adopt the perspective of developmental systems theory. Doing so allows us to preserve the insights of EP, whilst avoiding the temptation of accepting insufficiently detailed aetiologies. Before making that claim, however, I must first detail the objections which motivate it. It is to this task that I now turn.

(i) Evolutionary Just-So Stories

The most well-known of objections to EP is also the most general. It was first formulated by Stephen Jay Gould as a criticism of Wilson's sociobiology programme. Rather than challenging any of the specific claims made by Wilson, Gould questioned the legitimacy of looking for evolutionarily adaptive explanations of complex traits. Not all such traits, he argued, are helpfully conceived of as adaptations. Many may have arisen as a result of selection for *other* traits. Furthermore, it is likely that it will typically be possible to develop *some form of* adaptive explanation for any existing trait, behavioural or otherwise. This is because if such traits were highly *mal*adaptive, they would have been selected against and would therefore no longer exist. Still, Gould argues, the fact that it is *logically possible* to provide an adaptive explanation for a trait does not in any way guarantee the *truth* of that explanation. For example, Gould questions the plausibility of adaptive explanations of social co-operation and conformity.

> Societies without cohesion may not survive, even if their members have the same genetic make-up as others who learned to cooperate and transmitted these beneficial habits culturally.[67] Cooperation and conformity are within the capacity of the human genome; they need not be coded as specific adaptations.[68]

The same concern has been expressed more recently by Francisco Ayala, who writes that:

> [m]oral codes, like any other dimensions of cultural systems, depend on the existence of human biological nature and must be consistent with it in the sense that they could not counteract it without promoting their own demise. [...] Discrepancies between

[67] It is important to note that when Gould mentions cultural transmission in this quotation, he does not have in mind the heuristic selection account of cultural transmission discussed in §4.ii. Rather, he means unbiased social learning.

[68] Gould (1980) p. 287.

> accepted moral rules and biological survival are [...] necessarily limited in scope or would otherwise lead to the extinction of the groups accepting such discrepant rules.[69]

Gould and Ayala certainly raise an important concern. However, this worry is not a knock-down objection to EP. Thus, Richerson and Boyd concede that anti-adaptationists:

> are surely right that we should be cautious about accepting adaptive "just-so" stories about the function of traits that we observe. But we should be equally cautious, perhaps more cautious, about casually accepting non-adaptive just-so stories that invoke mysterious unspecified events or tradeoffs.[70]

That is to say, uncritically accepting a non-adaptive explanation for a trait with a clearly defined function makes accounting for the existence of that trait incredibly difficult. Explaining how that trait evolved then becomes a matter of speculation regarding how it might have been non-selectively produced as a by-product of other traits. Furthermore, adaptive explanations have an additional advantage over non-adaptive explanations. As Tooby and Cosmides explain:

> adaptationist approaches offer the explanation for why the psychic unity of mankind is genuine and not just an ideological fiction; for why it applies in a privileged way to the most significant, global, functional, and complexly organised dimensions of our architecture [...]. If the anti-adaptationists were correct [...] and our evolved architectures were not predominantly sets of complex adaptations or properties developmentally coupled to them, then selection would not act to impose cross-individual uniformity, and individuals would be free to vary in important ways and to any degree from other humans due to genetic differences.[71]

Of course, it does not follow from the responses made by adaptationists that any adaptive explanation of a trait *whatsoever* is likely to be the correct one. There may be competing adaptive explanations, and the possibility that some traits are non-adaptive by-products of the sort envisaged by Gould cannot altogether be excluded. Nevertheless, adaptive explanations are the best starting point for conceptualising the evolutionary history of complex traits.

(ii) Are Pleistocene Minds Maladaptive in Modern Societies?

It was noted in §3 that a typical feature of EP's conception of the human mind is its view that evolved cognitive adaptations are tailored to meet the requirements of a Pleistocene environment. According to this view, it will be recalled, the human mind has been millions of years in the making, and evolution proceeds at far too slow a pace for significant adaptive changes to have been made since the advent of the earliest agricultural communities. Given that our environment has changed so profoundly in so short a space of evolutionary time, evolutionary psychologists argue that a

---

[69] Ayala (2006) p. 149.
[70] Richerson and Boyd (2005) p. 103.
[71] Tooby and Cosmides (1992) p. 79.

number of our cognitive adaptations will prove *maladaptive* in modern societies (e.g. an innate preference for foods high in sugar).

This assumption, termed the "adaptive-lag hypothesis",[72] has been called into question by research into human behavioural ecology. This research makes use of the concept of "niche construction". From a niche construction perspective, organisms are seen not merely as passively inhabiting their environment, whilst being acted upon by the forces of natural selection. Rather, they are seen as actively re-structuring that environment in order to make it more hospitable; and as changing the way in which natural selection acts upon them in the process. That niche construction takes place is uncontroversial. Examples include bird's nests, beaver's dams, and changes in soil chemical levels induced by plants.[73] What is more controversial, however, is the role which niche construction plays in evolution: "standard evolutionary theory does not deny niche construction, but interprets it solely as a product of evolution rather than as part of the process".[74]

Human beings are, according to Kevin Laland and Gillian Brown, prolific niche constructors. This is a consequence of our uniquely developed capacity for the cultural transmission of information and tradition: "human cultural processes are exceptionally potent compared to those in other animals, probably because of [their] cumulative property".[75] It is this capacity for extensive niche construction that allows humans to live in such disparate environments.

What relevance, however, does the phenomenon of niche construction have for EP's adaptive-lag hypothesis? Put simply, niche construction allows humans to create environments which are in accord with, rather than opposed to, their evolved, adaptive dispositions.

> Like the acorn-storing squirrel or the wasp that cools her nest with droplets of water, our ancestors ensured the availability of food by tracking game and storing food, and controlled temperature by manufacturing clothes and building fires and shelters. In principle, modern refrigerators and air-conditioning are no different. [...] Human-built environments might be different from African savannah, but many selection pressures acting on us could be broadly similar, since our constructions were built to be suited to our bodies and their needs.[76]

Of course, niche-construction designed to secure foreseeable, short-term benefits may also bring about unforeseeable, long-term costs. Niche construction is can therefore be a double-edged sword. Yet in the event that certain unforeseen costs of niche construction *do* begin to manifest themselves, further cultural evolution creates the possibility of negating, or at least minimising, those costs. For example, the formation of large, sedentary communities facilitated the development of agriculture, and the nutritional benefits that come with having a regular, sustainable source of food. Such societies also exposed their members to a new source of selection pressure, in the form of "a host of diseases, including measles, smallpox, and typhoid, that thrive in dense populations with poor sanitation".[77] Cultural evolution eventually allowed these large communities

---

[72] Laland and Brown (2006) p. 98.
[73] See Laland and Brown (2006) p. 95 for further examples.
[74] Laland and Brown (2006) p. 95.
[75] Laland and Brown (2006) p. 96.
[76] Laland and Brown (2006) p. 99.
[77] Laland and Brown (2006) p. 100.

to mitigate these costs by developing solutions ranging "from sewerage plants to drains to water purification treatments".[78]

In the event that cultural innovations are unable to overcome this new selection pressure, that pressure will make genetic adaptation to the changed environment more likely. Laland and Brown argue that EP underestimates the potential rapidity of genetic evolution, and claim that "there are several examples of culturally induced genetic responses to human agriculture".[79] One such example is to be found in the evolution of an increased frequency of sickle-cell *S* alleles among West African yam farmers.

> These people cut clearings in forests to grow crops [...]. The clearings increased the amount of standing water, which provided better breeding grounds for mosquitos and increased the prevalence of malaria. This, in turn, modified natural selection pressures in favor of an increase in the frequency of the sickle-cell *S* allele because, in the heterozygous condition, the *S* allele confers protection against malaria. Here culture has not damped out natural selection, but rather induced it.[80]

Based on these considerations, human behavioural ecologists have questioned EP's claim that cultural evolution has created an atypically large adaptive lag between the human mind and modern societies. This debate is not one which I will attempt to resolve, as the issue of adaptive lag is not a part of the subject matter of this thesis. Nevertheless, the adaptive lag debate raises other considerations which *do* relate to the present discussion.

For the purposes of this chapter, the key issue raised by human behavioural ecology and the phenomenon of niche construction is the complex nature of the interaction between an organism and its environment. There is a reciprocally influential dynamic to this interaction which EP is insufficiently attentive to. Rather than seeing the development of a trait as the result of direct genetic selection for that trait, niche selection encourages us to see traits as the emergent products of an iterated series of interactions between both organism and environment. Organisms are not passively moulded by their environments. Rather, they actively shape them, simultaneously changing the selection pressures which influence the further evolution of their species. Adopting this perspective does not, of course, undermine the general strategy of EP. But it does suggest that its evolutionary narrative is insufficiently detailed. This is a point which will be returned to, and more fully developed, in §6.

(iii) Alternatives to Modular Explanations

Another worry about the adequacy of EP's narrative comes into view when alternative models of trait development are considered. According to EP, if a trait is both species-typical and invariant, then it is likely to be the product of an evolved mental module which was designed to

[78] Laland and Brown (2006) p. 100.
[79] Laland and Brown (2006) p. 101.
[80] Laland and Brown (2006) p. 101.

produce the trait in question. A simple example, which has already been briefly discussed, is a dietary preference for foods with high sugar content. The account which EP provides of this trait argues that such foods were hard to come by in the environment in which early humans' food preferences adapted. Individuals who preferred sugary foods to available alternatives got more energy from their diet, and so had a higher genetic fitness than their sugar avoiding conspecifics. This food preference was selected for, and is now an innate feature of the human mind's cognitive architecture. However, an alternate account of the development of this trait is available.

Eva Jablonka and Marion Lamb have argued that genetic selection is not the only process capable of generating evolutionary change. They identify three additional inheritance systems, each of which operates alongside, yet independently, of genetic selection. These are the "epigenetic" "behavioural" and "symbolic" inheritance systems. A discussion of each of these systems is unnecessary at this point. A short discussion of one of the outcomes of the behavioural inheritance system will suffice to show how a modular explanation of a trait can be called into question.

The behavioural inheritance system shows how seemingly modular food preferences may develop without genetic selection for those preferences taking place. To that end, Jablonka and Lamb discuss research undertaken on food preference formation in young rabbits. Such research shows that food preferences begin to form *in utero*, but these are not an innate (i.e. developmentally canalised) part of rabbit development. Food preferences are in fact conditioned by the contents of the mother's diet. Rabbits are born with a preference for those foods which their mothers consumed during pregnancy.

> The offspring of juniper-eating mothers had acquired information about juniper food from her while in the womb, presumably because chemical cues had reached them through the amniotic fluid and placenta.[81]

Further studies have shown that food preferences are also transmitted through maternal milk during weaning.[82] There is also evidence to suggest that this process is similarly influential in the formation of food preferences in human infants.

> Recently it has been found that the six-month old babies of women who had had a lot of carrot juice during the last three months of pregnancy preferred cereal made with carrot juice to that made with water. The same was true if the babies' mothers had had the carrot juice only during the first two months of the breastfeeding period. Babies whose mothers had drunk just water showed no such preference.[83]

There is good evolutionary reason for food preferences to develop in this way. If a pregnant mother is consuming a lot of a particular food, then it is highly likely that that food is not harmful. The mother's diet thus provides important clues to her developing infant about which of the current environment's food sources are safe for consumption. What this shows, is that a dietary preference can develop without genetic selection. Furthermore, if an infant retains the food preferences with which it is born, then it will pass those preferences on to its own offspring. A stable set of food preferences will therefore disperse throughout a lineage, and perhaps throughout an entire local

[81] Jablonka and Lamb (2005) p. 163.
[82] See Jablonka and Lamb (2005) p. 163.
[83] Jablonka and Lamb (2005) p. 163.

population, if the food in question is in more plentiful supply than possible alternatives. This is because the members of that lineage will have a better chance of finding their primary food source than conspecifics with different preferences. Behavioural inheritance systems are therefore able to explain some of the same phenomena as EP's modular theory of mind, though without doing so in terms of genetic evolution.

Jablonka and Lamb argue that evidence in favour of a modular explanation of a trait is easily overstated. It does not follow from the universality of a trait that that trait is modularly encoded. "An alternative possibility is that no one has identified the initial conditions that cause the behavior's apparent invariance because we *all* experience the conditions."[84] For example, it was originally believed that ducklings were able to identify correctly the calls of the species to which they belong, and that this was a result of a genetically hardwired module. However, it has been shown that this is not always the case. There is at least one species of duck which learns its species-typical call, but does so prior to hatching.

> It seems that during the development of the vocal system, while they are still in the egg, the birds exercise their sound-making apparatus, and consequently hear their own self-produced vocalisations.[85]

Another potential avenue for non-modular explanations is cultural evolution. As seen in §4 evolutionary psychologists *are* concerned with detailing the interaction between genes and culture. Yet Jablonka and Lamb argue that evolutionary psychologists typically "fail to recognize the power and subtlety of cultural evolution".[86]

> For example, think how difficult it was 1200 years ago for someone in Europe to divide one number by another. Say they wanted to divide 3712 by 116, or as it would have been back then, MMMDCCXII by CXVI. Using the Roman notation system, they would have needed an abacus or a set of tables to accomplish this task [...]. Today, with our Arabic notation system (and the useful zero), it takes the average ten-year-old only minutes to get the answer 32. If we knew nothing of the cultural change in the number notation systems and judged simply by the ability to learn to do sums quickly and correctly with just pen and paper, we might well deduce that during the last 1200 years a great mathematical mutation had occurred and been incorporated into our maths module through natural selection.[87]

Again, such criticisms cannot be said to show that there are *no* mental modules. Nor do they show that, in the absence of a more satisfying alternative, a modular explanation of a trait is not the correct one. Rather, what this criticism does suggest is that EP's reliance on modules as an explanatory device is far too simplistic. There are many alternative explanations for the existence of seemingly innate traits which EP systematically overlooks.

The objections to EP which this section has considered have not issued in a knock-down refutation. For all that has been said, the accounts which EP gives of incest avoidance, social

---

[84] Jablonka and Lamb (2005) p. 217.
[85] Jablonka and Lamb (2005) p. 218.
[86] Jablonka and Lamb (2005) p. 213.
[87] Jablonka and Lamb (2005) pp. 218 – 219.

reasoning, and the influence of our primary values on cultural evolution may all be true. Nevertheless, those objections do give reasons to be wary of EP's readiness to posit modular explanations for traits, and its tendency to overlook the complex interactive processes which produce those traits. In the following section, I will discuss a different approach that has been taken towards the conceptualisation of evolved traits, namely that of developmental systems theory. This approach is not susceptible to the criticisms which have been made of EP, yet it allows us to continue to posit and test the adaptive explanations which make EP such a useful explanatory device.

6. Developmental Systems Theory

At the heart of developmental systems theory (DST) there is a fundamental criticism, not only of EP, but also of sociobiology and of models of gene-culture coevolution like those developed by Durham and by Richerson and Boyd. To present this criticism to its fullest effect, it will be helpful briefly to recap the general narrative shared by each of the positions so far discussed.

Sociobiology, evolutionary psychology, and models of gene-culture coevolution hold that human behaviour is influenced by genetically programmed psychological tendencies. These tendencies have been naturally selected for over the course of evolutionary time. They are tendencies which disposed our ancestors to avoid genetically destructive actions such as incest, while also better equipping them to engage in fitness-increasing social interactions. Inevitably, however, there is far more to the complexities of social behaviour than the influence exerted upon us by our genes. The psychological dispositions with which natural selection has equipped us are very basic, and as such are capable of being expressed very differently, depending on the cultural context in which their possessors find themselves. Some of these contexts may even prevent these dispositions from developing into traits at all. For example, individuals living in an environment which does not contain snakes or spiders, nor any information about them, will not develop phobias of these creatures, no matter how readily disposed to do so they may be. It is therefore necessary to draw a distinction between the dispositions instilled in us by natural selection, and the environment in which those dispositions are expressed. The dispositions are universal, species-typical facts about human psychology. The environments in which they are expressed differ from one another dramatically, and constitute the source of cultural variation which we find between human societies.

This, very broadly, is the starting point shared by each of the theories so far discussed. This conceptual framework allows for a neat division to be drawn between the fixed, innate contents of human psychology on the one hand, and on the other, those aspects of psychology which are learned during the process of enculturation, and which are therefore malleable. This perspective can be easily coupled with the most widely accepted definition of evolution, according to which: "[e]volution is best understood as the genetic turnover of the individuals of every population from generation to generation".[88] This coupling allows for a fairly straightforward answer to the question

[88] Mayr (2001) p. 84. Italics removed.

"did morality evolve?" That answer is "yes, insofar as our moral beliefs are, to a certain extent, shaped by psychological dispositions coded for in our genes".

DST rejects this narrative. Its reason for so doing is that such a perspective ignores important ontogenetic details; i.e. facts about the way in which traits develop within particular organisms (as opposed to their development within the species to which those organisms belong). The alternative narrative which DST recommends does not only affect the way we think about development, however. It also suggests an entirely new way of thinking about evolution itself.

(i) Two Types of Interactionism

It is platitudinous to assert that the dichotomy between nature and nurture is a false one. The question of whether any given trait is a product of nature or nurture must always be answered with "both". This is something which each of the positions so far discussed agrees upon. Thus it was noted in §2.1 that, according to Wilson, "[e]ach person is molded by an interaction of his environment, especially his cultural environment, with the genes that affect social behavior."[89] However, according to adherents of DST, the sort of interaction which Wilson and those who follow in his tradition have in mind actually *preserves* the dichotomy which it purports to abolish.

Susan Oyama[90] highlights this issue by distinguishing between what she terms "standard interactionism" and "constructive interactionism". Each of the evolutionary interpretations of human behaviour so far considered is of the standard interactionism type. Oyama is highly critical of standard interactionism, and argues for the acceptance of constructive interactionism in its stead. How, then, does constructive interactionism differ from its standard rival?

According to Oyama, standard interactionism is committed to the view that the causal mechanisms which produce a trait can be hierarchically ordered. From the perspective of standard interactionism, genetic information is analogous to a blueprint for a particular organism. Whilst it is acknowledged that various environmental resources are necessary for the organism to develop in a species-typical way, these resources are given a secondary role. They are the raw materials with which the developmental master-plan contained in the genes is enacted. Thus the genes can be said to contain the essential biological truth about the organism. Thus she argues that from a standard interactionism perspective:

> DNA is given a special role in accounts of ontogenesis. The causal privileging in such accounts is not a matter of saying that only genes are needed for development (no one says this). It is the attribution of special directive, formative, or informative power to genes – in short, the treating of some causes as more equal than others.[91]

This view is intuitively attractive because genes are typically thought to be the only things capable of being inherited over the course of a species' evolutionary history. However,

---

[89] Wilson (1978) p. 18.
[90] Oyama (2000a) pp. S332 – S333.
[91] Oyama (2001) pp. 177 – 178.

developmental systems theorists point out that this is not the case. The environments in which development takes place are *also* inherited from one generation to the next, and the ways in which genes are expressed are necessarily influenced by the environment *in which* they are expressed. Constructive interactionism differs from standard interactionism by paying closer explanatory attention to, and emphasising the importance of, the role of environmental influences in trait development.

The environment in which an organism develops plays an extremely important role in the ontogenetic emergence of traits. Some aspects of developmental environments are relatively unchanging. This can be the result either of niche-constructing activity through which an organism sustains desirable features of its environment, or of the inherent immutability of certain types of environmental properties. These are therefore reliable sources of developmental information, which evolution can exploit in the development of an organism's traits. It does this by using information reliably present in the environment to regulate the expression of the genes. Thus, an appropriate environmental influence is just as essential for the development of traits traditionally considered innate as it is for those which are unique to an individual organism. Thus Oyama writes that:

> [t]here is a recurrent tendency to associate predictability and constancy mainly with insides, and (certain kinds of) change and variation with outsides. [...] But developmental constancy is no less a product of systematic interaction than is variation. In like manner, lability, unpredictability, and variability [are] no less (interactively) systematic than is constancy.[92]

This systematic, constructive interaction is ubiquitous. For example, the Barker effect shows that the development in adults of conditions such as diabetes and heart disease can be influenced by the environmental conditions to which that adult was exposed during foetal development. The Barker effect is caused by maternal under-nutrition, or rather the presence of placenta-crossing hormones such as cortisol, which accompany such under-nutrition. These hormones effectively act as a signalling system to the developing foetus, indicating that the environment into which it is about to be born does not contain an abundant supply of nutrition. As Melvin Konner describes it, these hormones instruct the foetus to develop "a streamlined, smaller version of the human body designed to survive and adapt under sparse nutrition".[93]

These hormonal signals are not always reliable, however. For instance, they may be the result of an atypical restriction in the mother's nutritional intake. Should an individual with such a "streamlined" phenotype be born into an environment which contains plentiful resources, they will effectively over-eat simply by consuming statistically normal amounts of food. This makes such individuals particularly vulnerable to the health problems associated with overeating. As Konner puts it, such individuals are "ideally adapted to sparse and stressful conditions but poorly prepared for abundant modern environments".[94] A rather harrowing confirmation of the Barker effect was discovered in the 1960s, as a result of experiences during the Second World War. During the German occupation of Holland:

---

[92] Oyama (2001) p. 188.
[93] Konner (2010) p. 540.
[94] Konner (2010) pp. 540 – 541.

> Dutch railway workers [...] called a strike to try to prevent German reinforcements reaching Arnhem. In retaliation, [...] [there was ordered] an embargo on all civilian transport in the country. The result was a devastating famine, which lasted for seven months. [...] [Researchers in the 1960s] found that those babies who were in their last trimester of gestation (only) [during the famine] suffered from low birth weight. These babies grew up normal but they later suffered from diabetes, probably bought on by the mismatch between their thrifty phenotype and the abundant rich food of the post-war world.[95]

When discussing the traits associated with the Barker effect, it is misguided to attempt to assign primary causal responsibility either to the genetic or to the environmental influences which go into producing them. On the one hand, there is a genetic predisposition toward metabolising food more slowly in order to conserve scarce energy resources. On the other hand, there is an environment in which those resources are amply present and which make the genetic tendency maladaptive. Moreover, the genetic predisposition is itself the result of a specific environmental stimulus, received whilst *in utero*.

The fact that the operation of the Barker effect can produce adaptive phenotypes in environments where food is often scarce suggests that it is an adaptive response to harsh environments. Here, then, is an example of selection working to develop traits through the mediation of environmental cues. This is precisely what constructive interactionism tells us to expect.

The ontogenetic dependency of adaptive traits on the presence of specific environmental stimuli may seem strange. If a trait is an adaptive one, why would natural selection allow its development to depend upon contingent, extra-genetic factors? After all, environments change, sometimes quickly. An organism which finds itself in an environment suddenly lacking certain features upon which its normal development depends will be much worse off than one whose traits can develop without that environmental input. Why, then, does development rely so heavily on the environment?

The answer to this question is simple. Environmental effects on development negate any selection pressure to develop adaptive traits in the absence of those same environmental effects. Selection would be unable to discriminate between traits which developed as a result of environmental stimuli, and identical traits which differed only in that they developed independently of environmental stimuli.

> The visual system of kittens does not develop normally without an input of patterned light [...]. Why, we could ask, didn't natural selection give them a *really* good visual system, one that would reliably grow with or without patterned light? The reason is that, since kittens have never been called upon to grow such a system, there was no disadvantage for those that could not.[96]

---

[95] Ridley (2003) p. 156.
[96] Konner (2010) p. 344.

(ii) Reconceiving Evolution

As noted above, the traditional conception of evolution, accepted by sociobiology and EP, has it that the term "evolution" refers to a change in gene frequencies within a population, observed over inter-generational time. Based on her account of the shortcomings of standard interactionism, Oyama argues that the traditional conception is not fit for purpose.

> [If, as argued] genes do not create traits according to a plan written in their very structure, [...] if phenotypic characteristics arise only when sufficient interactants are present in the proper place and at the proper time, and if all these factors are therefore given comparable causal and formative significance, *then defining heredity as the passing on of all developmental conditions, in whatever manner,* is preferable to defining it by genetic information.[97]

Oyama therefore proposes that the term "evolution" should be used to refer to "the change in constitution and distribution of developmental systems, organism-environment complexes that change over both ontogenetic and phylogenetic time.[98] This is the fundamental claim made by DST, and it is this which distinguishes its conceptual approach from that of the other theories discussed earlier in this chapter.

As might be expected, accepting DST has important repercussions for how we think about human traits. One of the most important features of the human developmental system is of course culture. As Paul Griffiths and Russell Gray note, "humans have had a culture since before they were human".[99] As a result, "[m]any species-typical features of human psychology may depend critically on stably replicated features of human culture".[100] That is, there may be many traits which are universally acquired, but which we would not develop in the absence of specific cultural inputs. Language is perhaps the most striking example of such a trait. As Steven Pinker and Paul Bloom note:

> [a]ll human societies have language. As far as we know, they always did [...]. [T]he ability to use a natural language belongs more to the study of human biology than human culture.[101]

Notwithstanding Pinker and Bloom's claims, however, a child raised outside of a community of language users will not develop language skills. Cultural input is essential for the ontogenesis of linguistic traits.

Griffiths and Gray agree with Oyama's assessment that "evolution is best construed as differential replication of total developmental processes or life cycles".[102] The importance which human cultural systems play in those life cycles lead Griffiths and Gray to claim that:

---

[97] Oyama (1985) p. 43.
[98] Oyama (2000b) p. 148
[99] Griffiths and Gray (1994) p. 302.
[100] Griffiths and Gray (1994) p. 302.
[101] Pinker and Bloom (1992) p. 451.
[102] Griffiths and Gray (1994) p. 278.

> the developmental systems view makes it impossible to maintain the distinction between biological and cultural evolution.[103]

As might be expected, accepting the developmental systems account of evolution therefore also means accepting a more inclusive enumeration of those traits which we think of as being the product of evolution.

> Electric light sockets have as yet played little role in human evolutionary history, yet my fear of them has an evolutionary explanation. The key lies in choosing the right description. Fear of objects associated with injury, or with fear displays in conspecifics is an evolved developmental outcome. There are evolutionary explanations of my acquiring a fear of any such object. So the resources that produce an organism with such fears are parts of the developmental system.[104]

This does not mean that according to DST, all conceivable traits are to be described as having evolved. Only those traits which are the products of a stably recurring developmental system can be said to have done so. Traits which are contingent upon the particular life history of an individual, which Griffiths and Gray term "individual traits", are therefore *not* the products of evolution. They explain this by way of an example of tissue scarring.

> There is an evolutionary explanation of the fact that the authors of this paper have a thumb on each hand. We have thumbs because of the replication of thumbed ancestors. The thumb is an evolved trait. But the fact that one of us has a scar on his left hand has no such explanation. The scar is an individual trait [...]. The resources that produced the thumb are part of the developmental system. Some of those which produced the scar, such as the surgeon's knife, are not.[105]

DST purports, then, to provide a more accurate way of conceptualising evolution than the traditional view. Its claims to this effect may seem highly controversial, given the association of evolution with genes. There are two important points to bear in mind when considering this issue, however. The first is that evolutionary theory antedates the discovery of the gene. The deep association between these two concepts was a result of the emergence of the Modern Synthesis during the 1930s. The term "gene" is therefore not an essential conceptual part of the term "evolution". The second point worthy of note is that DST's reconceived account of evolution by no means ignores genes. An organism's genetic code is an essential part of its developmental system, and changes to this code will sometimes result in changes to that organism's phenotypic traits. DST is not a theory which ignores this influence. It simply reconciles it with the equally important influence of the organism's ontogenetic environment.

When these points are taken into consideration, they ought to dispel any concerns that DST in some way violates an incontrovertible fact about what evolution "really is". This now having been made clear, the attractions of adopting a developmental systems perspective become compelling. DST provides a more accurate account of the development of traits, and of how development depends upon particular, predictably recurring environmental stimuli. In addition to this, however, it

---

[103] Griffiths and Gray (1994) pp. 278 – 279.
[104] Griffiths and Gray (1994) p. 297.
[105] Griffiths and Gray (1994) p. 286.

is also not susceptible to the criticisms made of EP's approach, discussed in §5. In the next section, I will briefly explain why this is the case.

(iii) DST and Constructive Interactionism are Preferable Alternatives to EP and Standard Interactionism

One of the criticisms of EP discussed above was that its usage of adaptive explanations is over-zealous. By its extension of the scope of evolutionary explanations to include all aspects of a developmental system, it may appear that DST will be even more open to this objection than was EP. Yet this is not the case. Culturally developed moral systems *have* evolved, according to DST, as they constitute a part of a stable, and therefore heritable, developmental system. Yet it is not accurate to infer from this that such systems have been adaptively selected for. As Griffiths and Gray explain:

> [e]volutionary explanation is "adaptive-historical" explanation. The organism's response to any particular adaptive phase is determined in part by the historical resources and historical constraints accumulated in the lineage in response to past phases. [...] The outcome is also influenced by the availability and order of variants and by the sheer stochasticity of the differential replication process. So even in cases where adaptation plays a role in the explanation of a particular trait, that explanation is very far from adaptationist.[106]

DST therefore allows for the adaptive analysis of traits, but does not begin such analyses with the assumption that adaptive explanations are necessarily the *correct* explanations. This should make it a less objectionable analytical tool for those sympathetic to Gould and Ayala's scepticism of EP.

Another objection raised against EP was that made by human behavioural ecologists, who argued via niche construction theory that EP tends to neglect the complexities of organism-environment interaction. It is important to recall that this criticism was made from within niche construction theory's standard interactionism perspective, and is therefore conceptually distinct from Oyama's constructive interactionism. This difference has been highlighted by Kim Sterelny, who argues that niche construction theorists:

> have some tendency to focus on just two streams of cross-generation influence: genetic inheritance and cultural inheritance, with genetic inheritance being in some way primary, but modified by a secondary system. Other factors are important because they modify, and are modified by, the genetic inheritance system.[107]

Nevertheless, DST and niche construction theory do share similar concerns. Thus they are "the only two approaches that identify an intrinsic connection between environmental engineering and a critical role for nongenetic inheritance".[108] This similarity has lead Sterelny to describe niche

---

[106] Griffiths and Gray (1994) p. 287.
[107] Sterelny (2001) p. 336.
[108] Sterelny (2001) p. 337.

construction theory as DST's "conservative cousin".[109] The concerns shared by these approaches mean that DST is not susceptible to the criticism which niche construction theorists make of EP's insufficiently interactive conception of evolution.

Finally, it was seen that EP's reliance on modular explanations of complex behavioural traits is problematic. Such emphasis ignores the possibility of non-modular explanations which are just as able to account for the same traits. It should already be clear that this is not an issue for DST. DST makes no predictions about the modular basis of a trait: its approach simply emphasises the need to pay close attention to the resources, both environmental and genetic, that are required for a trait to develop. Note, however, that DST does not begin with the assumption that such traits must be, or are likely to be, *non*-modular. Developmental systems theorists can remain officially agnostic as regards this issue, resolving it only on a trait-by-trait basis.

For each of these reasons, then, DST presents a more attractive theoretical framework than EP and other standard interactionism approaches.

7. Conclusion

In this chapter I have discussed how evolutionary data can be used to interpret human behaviour. §2 dealt with sociobiology, the influential starting point of modern approaches to the topic. There it was seen that evolutionary explanations of human behaviour do not entail controversial statements about the genetic determination of human beliefs or cultural practices. Rather, the way in which attitudes and opinions are formed can be seen as subject to the influence of learning biases. These incline an individual towards the development of certain psychological and behavioural traits, and away from others. These learning biases are formed as a result of the operation of epigenetic rules: genetically coded developmental programmes involved in the structuring of the brain, and the emergence of consciousness.

§3 introduced the research programme of evolutionary psychology. It highlighted the differences between this approach and that of sociobiology. The most significant of these is EP's interpretation of much complex behaviour as the result of naturally selected mental modules. The reasoning behind this interpretation, and of how it can be used to account for behaviour, was illustrated via a discussion of Cosmides and Tooby's analysis of social contract reasoning.

Following this discussion, §4 explained how the operation of basic, species-typical cognitive biases may nevertheless generate, and help to explain, significant amounts of cultural variation. This was shown, firstly, by an account of Durham's distinction between primary and secondary values. Genetically evolved primary values exert a consistent affective influence upon the assessment of mutually exclusive allomemes, making the adoption of certain cultural practices intuitively more desirable than others. Secondly, it was explained that cultural selection (i.e. selection between competing allomemes) is guided by the operation of selective heuristics. The operation of these heuristics guides cultural evolution towards genetically adaptive practices. However, the same

[109] Sterelny (2001) p. 337.

heuristics, when operating in diverse environments, are capable of generating significant amounts of cultural variation.

With a general outline of how evolutionary theory relates to the study of human behaviour and cultural change having been set out, §5 discussed some of the criticisms which have been made of that approach. It was seen that, whilst there are no knock-down objections to the evolutionary analysis of human behaviour, EP does have some shortcomings. The most significant of these was its failure to pay close enough attention to the role played by the environment during ontogenesis. This neglect on the part of EP means that the narrative which it advocates is too simplistic, and ignores the possibility of extra-genetic, non-modular explanations of the traits which it studies. This is a serious weakness in its explanatory methodology.

Finally, in §6, DST was introduced as a means of retaining an evolutionary perspective on human behaviour whilst avoiding the shortcomings of EP. DST's criticism of standard interactionism and its recommendation of an alternative, constructive interactionism were both set out. It was then shown that this alternative not only provides a more accurate ontogenetic narrative, but also avoids the difficulties associated with EP.

In the next chapter, the focus on DST will be retained. There, I will argue that by approaching evolutionary theories of ethics from a developmental systems perspective, much confusion over what is to count as an evolutionary ethical theory can be cleared up. This confusion has in fact stemmed from a tacit acceptance of standard conceptions of interactionism. By accepting Oyama's alternative, constructive account, this confusion can be avoided. Just as accepting this account has repercussions regarding how we think about evolution, however, it has similar repercussions regarding how we ought to understand claims that morality has evolved.

Chapter Two

The Scope of Evolutionary Theories of Ethics

1. Introduction

In this chapter, I consider the implications of constructive interactionism for the debate surrounding the evolutionary status of morality. Can the claim that morality has evolved be plausibly resisted, and if not, does this mean that we should all become *de facto* evolutionary ethicists? In §2 I show that both evolutionary and non-evolutionary interpretations of morality proceed from the shared perspective of standard interactionism. Despite this shared perspective, it seems *prima facie* plausible that there must be an empirical point of disagreement between evolutionary and non-evolutionary interpretations of morality, and that this disagreement is what allows claims about the evolutionary status of morality to be rejected. As will be seen in §3 however, this is not the case. Adopting the developmental systems perspective endorsed at the end of the previous chapter reveals there to be various confusions at work in denials of the evolutionary status of morality, not least of which is the degree to which non-evolutionary theories are empirically distinct from their evolutionary counterparts. Adopting constructive interactionism allows us to avoid these confusions, but it also shows non-evolutionary interpretations of morality to be deeply problematic.

Still, it does not follow from this that all ethical theories are necessarily evolutionary. In order to see why this is the case, it is helpful to point to a different way of distinguishing between evolutionary and non-evolutionary ethical theories. I do this in §4 by drawing attention to the difference between moral psychology and moral philosophy. Rather than attempting to defend the implausible position that moral development has not been influenced by evolution, non-evolutionary ethicists should deny that this influence has any normative or metaethical significance. Whilst both of these denials are philosophically possible, I argue that many metaethical positions ultimately draw on evolutionary data, and must therefore be reclassified as evolutionary.

To the extent that this conclusion is counterintuitive, it might be thought to constitute a *reductio* of the previous chapter's argument for adopting a developmental systems perspective. I briefly consider this objection in §5. Against it, I argue that any conceptual or terminological revision which follows from accepting constructive interactionism is virtuous, insofar as it allows us more easily to avoid perpetuating the confusion currently found in the literature. Furthermore, endorsing an evolutionary perspective on morality does not require one to accept specific normative or metaethical positions. The consequences of adopting a developmental systems perspective are therefore innocuous.

2. A Shared Interactionism

In the previous chapter, evolutionary interpretations of human behaviour were seen to endorse (either standard or constructive) interactionism. This is the claim that traits emerge as a result of interactions between our genes and our environment.[1] Interactionism is also espoused by (putatively) non-evolutionary interpretations of behaviour, and ethics is no exception to this rule. Because the literature on evolved mental traits has largely been dominated by evolutionary psychology, both evolutionary and non-evolutionary theories of ethics tend to proceed from an evolutionary psychological perspective. Both, therefore, tend to subscribe to standard models of interactionism. As argued in the previous chapter, standard interactionism is not the best way to conceptualise the development of traits.[2] As will be seen in the next section, an un-critical acceptance of standard interactionism has led to conceptual confusion in the literature on the evolutionary status of morality. Adopting a developmental systems perspective allows this confusion to be more easily avoided, and reveals the ubiquity of evolutionary explanations. But could a non-evolutionary theory of ethics avoid evolutionary data entirely, and simply reject interactionism? In fact it could not plausibly do so. That this is the case can be seen more clearly by describing what a non-interactive account of morality would have to look like.

A non-interactive *evolutionary* account of morality would subscribe to the implausible claim that morality resides in our genes alone. It would hold that moral beliefs develop solely as a result of our possessing certain genes, or gene complexes. No amount of environmental stimuli would have any influence on the moral codes which we individually develop, according to such a theory. As seen in the previous chapter, no evolutionary ethicists actually make this claim: the account it proposes is simply biologically false.

However, a non-interactive, *non*-evolutionary account of morality is equally implausible. Such an approach would be committed to the claim that the development of our moral beliefs has nothing whatsoever to do with our genes. Yet this claim is clearly too strong. Denying the role of a genetic factor in moral development would render such a theory unable to account for the fact that morality does *not* develop in non-human species. Even if it could be argued that *some* non-human animals, for example other primates, can in fact make moral judgements (a claim which is typically denied[3]), it is uncontroversial that the vast majority of species do not. Neither, for that matter, do plants or bacteria. The point is of course an obvious one, but to explain why some organisms make moral judgements whilst others do not is impossible without saying something biologically substantive about the kinds of organism in question. Now, it is possible to say something about organisms *qua* species without doing so explicitly in terms of genes. Nevertheless, the genes possessed by those organisms are a part of what makes them what they are. Thus genetic differences are an essential part of any full explanation of why some organisms do not develop morality.

Accounting for the species-typical differences between organisms which prevent non-human species from developing morality therefore leads to a perspective on morality that is at least minimally interactive. Yet there is much more that can be said about the kind of creatures which human beings are, and which has been thought relevant to the study of morality. For example, virtue ethicists make claims about the kind of activities which make human lives go well; and

---

[1] Here, "environment" should be read as inclusive of the cultural practices into which we are born.
[2] See Chapter One §6.
[3] See de Waal (2006) for discussion.

hedonists talk about human beings' capacities to experience pleasure and pain, and the intrinsic goodness and badness of those respective states. Such claims are typically supported by statements about the kind of creatures that human beings are. Given that we have evolved to be such creatures, there is an essentially genetic component to any fully-spelled-out account of what makes a human life a good one.

Any philosophically respectable theory of the development of morality is therefore committed to an at least minimally interactive account of morality; furthermore, such theories will often be somewhat more than minimally interactive. Thus, both evolutionary and non-evolutionary ethical theories take morality to develop out of the interaction between our genes and our environment. In what respect, then, do these theories differ?

3. How Not to Reject Evolutionary Theories of Ethics

There are three ways of denying the evolutionary status of morality which seem, *prima facie*, more plausible than denying the minimally interactive account described above. The first is to downplay the causal importance of the role played by the genes in the formation of moral beliefs. The second is to deny that morality is a naturally selected adaptation. Both of these strategies can be found in the work of Jesse J. Prinz. In the following two sub-sections I will focus on his presentation of each of these approaches. The criticisms which I make of Prinz's account do not apply only to his position, being sufficiently general to extend to other anti-evolutionary accounts of the same stripe.

A third way of denying that morality has evolved is to deny that it is the product of an evolved developmental system. To the best of my knowledge, this is not an approach which has been advocated in the literature. This is no doubt due to an absence of discussion of developmental systems theory in the literature on evolutionary ethics. Nevertheless, for the sake of inclusiveness I give a brief account of such an approach, and of the difficulties which it entails, in sub-section three.

As will be seen from the difficulties encountered by these approaches, rejecting the claim that morality has evolved is not as straightforward, or as uncontroversial, as it may at first seem.

(i) Genes are of Secondary Causal Importance in the Development of Morality

Prinz accepts that the non-evolutionary interpretation of ethics for which he argues proceeds from the perspective of interactionism. Thus he observes that "[i]t doesn't make sense to ask whether human beings are a product of nature *or* nurture. Obviously, the answer is both".[4] Rather, according to Prinz, the pivotal issue is one of causal priority. Thus he asks: "[a]re we, by nature, primarily driven by innate, evolved and genetically controlled traits, or are we primarily

[4] Prinz (2012) p. 6.

driven by experience?"[5] Clearly then, Prinz subscribes to standard interactionism, one of the defining features of which is the view that "some causes [are] more equal than others."[6]

For Prinz, the most important factor in development is experience, particularly when the developing trait in question is morality. It is this intuition which underlies his claim that "the wide range of moral rules found cross-culturally suggests that children can acquire moral attitudes towards just about anything".[7] Prinz is aware of the attempts, discussed in the previous chapter, to supplement evolutionary psychology made by evolutionary anthropologists such as William Durham, Peter Richerson, and Robert Boyd. Whilst he describes such work as providing "valuable contributions",[8] Prinz holds that they do not sufficiently emphasise the role played by the environment in the development of our moral beliefs.

> Scientific debates are rarely settled by compromise [...]. [I]f some traits are strongly influenced by culture, the involvement of genes is no more informative than the involvement of quarks – sure, we need genes to behave, but genes add little explanation of why we behave one way rather than another.[9]

Of particular relevance to moral traits is the influence exerted by our cultural environment: "evolutionary ethicists systematically underestimate the contributions of culture".[10]

Prinz's position on this issue is at odds with the developmental perspective endorsed in the previous chapter. According to developmental systems theorists, the question which Prinz sets himself the task of answering (i.e. "to what extent are our moral beliefs produced either by our genes or our environment?") is not one which can coherently be asked. Why, then, does Prinz think that it *can* be asked; and why, specifically, is he mistaken?

Prinz's error is attributable to a misconception, on his part, of heritability studies. Prinz claims that:

> heritability studies [...] estimate genetic contribution to behaviour by measuring correlations between traits and familial relatedness. If a trait runs in families, it's more likely to be genetic [...].[11]

However, this analysis is mistaken. This can be seen by looking at Evelyn Fox Keller's exposition of the term "heritability". According to Keller, when it is used by population geneticists "heritability" has a technical meaning unrelated to that of the non-technical term "heritable". Thus "heritability" refers to "a statistical quantity associated with the ratio of genetic variation to phenotypic variation within a specified population of organisms".[12] That is to say that the term "heritability":

---

[5] Prinz (2012) p. 13.
[6] Oyama (2001) pp. 178.
[7] Prinz (2008a) p. 164.
[8] Prinz (2012) p. 12.
[9] Prinz (2012) p. 12.
[10] Prinz (2007) p. 285.
[11] Prinz (2012) p. 18.
[12] Keller (2010) p. 57.

> has meaning only in relation to the properties of a population, not to properties either of an individual or of an individual lineage. [...] In other words, I can ask if my musical ability [...] is heritable, but I cannot ask, what is the heritability of my musical ability [...]?[13]

Furthermore, heritability studies take *differences* in traits as their explanandum, and *not* continuity of traits, as implied by Prinz's account. Accordingly, even traits that have a clearly genetic basis may exhibit very low scores in tests of genetic heritability.

> We would normally say that hand number is a heritable trait. But what is its technical heritability? Answer: zero, or very close to it. And the reason is that, while there is phenotypic variance in the human population (not everyone has two hands), this variance is almost entirely due to accidents, not to genetics. The genetic variance relevant to hand number in the population at large is virtually nil.[14]

Nevertheless, it might still be thought that heritability studies can tell us *something* about whether a trait is more or less the product of genes rather than the environment, or *vice versa*. If a trait is mutable under genetic variation, and less so under environmental variation, does that not suggest that genes are more causally responsible for the development of that trait? In fact it does not. Keller explains why this is the case by developing an analogy between gene-environment interaction in the production of traits, and drummer-drum interaction in the production of sound.

> It is useless to ask whether the drumming that we hear in the distance is made by the percussionist or his instrument because each of the two variables on which the sound [depends] [...] is influenced by the other in ways that do not permit separation. Yet if we hear two different sounds of drumming in the distance, we can ask and perhaps determine whether the difference between the two sounds is caused by a difference in drummers or by a difference in drums, or even how much of the net difference in sound is caused by the former and how much by the latter [...].[15]

Even in the event that such determination is possible, however, it would not thereby specify the extent to which either drummer or drum is causally responsible for the generation of sound in any single, particular instance. This question remains invalid.

> [C]alculated ratios of genetic variance to overall variance in phenotype are meaningless in the presence of either statistical interaction (when genetic and environmental variations are correlated) or constitutive interaction (when genetic and environmental effects are intertwined), for under such circumstances phenotypic variance cannot be partitioned into a genetic component plus an environmental component. This criticism is of particular relevance to human behavioral genetics for the simple reason that such interactions are ubiquitous in the development of human behavior. [...] Bluntly put,

[13] Keller (2010) pp. 58 – 59.
[14] Keller (2010) p. 61.
[15] Keller (2010) p. 35.

technical heritability neither depends on, nor implies anything about, the mechanisms of transmission (inheritance) from parent to offspring.[16]

Prinz's misconception of heritability studies, and of what they set out to show, underlies his belief that it is possible to meaningfully attribute a greater or lesser degree of environmental/genetic causation to the development of traits such as moral beliefs. Yet even if this *were* possible, it would be undesirable to draw the line between evolved/non-evolved in such a way. Just how *much* influence would genes have to have on development for a trait to be classifiable as the product of evolution? One hundred per cent? Eighty per cent? Or perhaps only fifty-one per cent? What then of a trait whose development is determined by fifty per cent genetic, and fifty per cent environmental influences? Ought such a trait be said to have evolved, or not? Thinking about evolution in terms of causal contribution makes our description of traits as evolved somewhat arbitrary, and this is an unacceptable consequence.

Thus the first strategy for denying the evolutionary status of morality, i.e. to claim that culture is more causally responsible for our moral beliefs than our genes, is incoherent. It rests on the mistaken assumption, inherent in standard interactionism, that the various causal influences responsible for the development of a trait are separable from one another. The developmental perspective of constructive interactionism allows us to see more clearly the conceptual problems which such an approach encounters.

With the first argument for a non-evolutionary interpretation of ethics having been addressed, it is now time to turn to the second such argument: the denial the morality is a naturally selected adaptation.

(ii) Morality is Not an Adaptation

The claim that morality is not a naturally selected adaptation constitutes a rejection of one of the central premises of both sociobiology and evolutionary psychology. If successful, then, it might seem to vindicate the claim that morality has not evolved. In fact, however, this is not the case. We freely, and legitimately, talk of non-adaptive traits as the product of evolution. Morality need be no exception to this practice. Before showing why it is innocuous to claim that non-adaptive traits have evolved, however, it will be helpful to outline Prinz's argument against construing morality as an adaptation. This argument targets the claim made by Leda Cosmides and John Tooby, discussed in the previous chapter, that there is an evolved mental module devoted to the regulation of social exchange.

Throughout his work on the (non-)relation of morality to evolution, Prinz conflates the terms "evolved", "innate", and "modular". He writes, for example, that:

> Evolutionary ethicists [...] tend to agree that morality is part of the bioprogram [...]. Recently, researchers have begun to look for moral modules in the brain, and they have

[16] Keller (2010) pp. 60 – 61.

> been increasingly tempted to speculate about [a] moral acquisition device and innate faculty for norm acquisition [...].[17]

Elsewhere, we read that "[i]f morality is innate that means there are specialized psychological mechanisms dedicated to thinking about morality".[18] And, finally:

> The open-endedness of morality suggests that the cultures in which we live actually contribute to the content of our moral rules, rather than selecting from a set of rules that are pre-coded in the genes. [...] If I am right, then it is a gross exaggeration to say that moral rules are products of evolution.[19]

This conflation is no doubt a consequence of the theoretical eclipse of sociobiology by evolutionary psychology. Prinz identifies the claim that morality evolved with the claims about modular adaptations made by evolutionary psychologists. As a result, he mistakenly identifies his argument against the modular basis of morality as an argument against the claim that morality has evolved *tout court*. With this caveat in mind, it is now time to consider Prinz's critique of Cosmides' and Tooby's argument for a social contract module.

The heart of the matter lies in the logical structure of Cosmides' and Tooby's version of the Wason selection task.[20] Prinz argues that when subjects were asked to test for violation of a non-contractual rule, the best strategy for successfully doing so was different from that in the scenario in which they were asked to test for the violation of a contractual rule. It will be recalled that Cosmides' and Tooby argue that in both contractual and non-contractual formulations, the correct answer to the Wason selection task is to check instances in which, when given in the form "if P then Q", P applies but Q is violated. Subjects ought not to test the conditional by checking cases in which Q applies, but they should check cases of non-Q to see whether it is the case that P.

Prinz describes the non-contractual, descriptive formulation of the Wason selection task as "a confirmation problem".[21] That is, the task asks subjects to look for evidence which supports the rule. In such cases, Prinz argues:

> [i]t is like trying to determine whether all ravens are black [i.e. If it is a raven (P) then it is black (Q)]. To do that efficiently, it's a good idea to check the ravens that we encounter, but a total waste of time to check all the things that aren't black. The logical fallacy of avoiding [cases in which Q is violated] is actually a good inductive strategy.[22]

Conversely, when the selection task is formulated as a social contract, "we assume the conditional is true; [...] rules are true [i.e. they are rules] even if they are not always followed".[23] Prinz argues that this shows that:

---

[17] Prinz (2008b) p. 367.
[18] Prinz (2012) p. 316.
[19] Prinz (2007) p. 286.
[20] See Chapter 1, §3.ii for discussion.
[21] Prinz (2007) p. 265.
[22] Prinz (2007) p. 265.
[23] Prinz (2007) pp. 265 – 266.

> there is an intrinsic difference between the task of assessing descriptive conditionals, in which case we are looking for counterexamples, and the task of assessing social exchange conditions, in which case we assume the conditional is true and look for violators.[24]

Based on this argument, Prinz rejects the claim that Cosmides' and Tooby's reformulation of the Wason selection task provides evidence for the existence of a cognitive module adapted for social contract reasoning. This is because subjects in the Wason selection experiment would not have approached the tasks which they were set in the same way. Their different performances on these tasks could therefore be attributed to subject's conceiving each task differently, and not to the existence of a mental module dedicated to reasoning about social exchange.

Prinz's criticism of Cosmides' and Tooby's argument for a social reasoning module is persuasive. It should be noted, however, that Cosmides' and Tooby's research also dealt with various reformulations of the Wason selection task, and that not all of these had the logical structure of "if P then Q".[25] Their findings were consistent across these reformulations, with subjects performing better on logical tests formulated as social contracts. It is unclear of the extent to which Prinz's criticism applies to each of these reformulations, and so his argument against the existence of a social contract module is by no means complete. For present purposes, however, I will assume that Prinz has successfully cast doubt upon the existence of a module dedicated to social exchange reasoning. Even if this much is granted, however, Prinz has by no means shown that morality is not the product of evolution.

If morality is not a modular adaptation, then what is it? The answer, according to Prinz, is that morality "is a by-product – accidental or invented – of faculties that evolved for other purposes".[26] Prinz suggests that these faculties may include memory, which allows us to maintain an interest in punishing moral transgressors even in the event that they evade detection for a protracted period of time; a capacity for imitation, which encourages the uniform punishment of similar transgressions; and the ability to empathise with others, which allows us to understand how our actions affect those around us.[27]

The literature on evolutionary biology has a term for traits which are not adaptations. They are called "exaptations". The term was introduced by Stephen Jay Gould and Elizabeth Vrba[28] as part of their critique of the adaptationist programme.[29] An exaptation is a trait which has not itself been selected for, but which supervenes on other adaptive traits which *were* the targets of selection. This is not to suggest that exaptations do not serve a purpose, however. They may be used by an organism to increase its genetic fitness in the same way that adaptations are so used. Exaptations differ from adaptations only in that they were not selected for because of their potential to increase fitness. Rather, they are the serendipitous by-products of selection for other traits. Nor are exaptations the only type of by-product to be postulated by Gould, who, with Richard Lewontin, also

---

[24] Prinz (2007) p. 266.
[25] See Cosmides and Tooby (1992).
[26] Prinz (2008b) p. 368.
[27] See Prinz (2007) p. 270 for discussion.
[28] In Gould and Vrba (1982).
[29] See Chapter 1 §5.i.

introduced the term "spandrel".[30] Like exaptations, spandrels are by-products. However, spandrels differ from exaptations in that they have no function, and therefore do not increase the fitness of an organism.[31]

Gould gives numerous examples of traits which he takes to be by-products. Of most relevance to the current discussion are complex cultural behaviours. Thus he argues that:

> [n]atural selection built the brain; yet, by virtue of structural complexities so engendered, the same brain can perform a plethora of tasks that may later become central to culture, but that are [by-products] rather than targets of the original selection – singing Wagner [...], not to mention reading and writing.[32]

Prinz clearly takes morality to be some form of exaptation, although he does not employ that term in his discussion. But to claim that a trait is an exaptation is by no means the same as to claim that it is not the product of evolution. Exaptations and spandrels are just as much the products of evolution as adaptations.

> Selection is responsible for producing the original adaptations that are then available for co-optation. It is responsible for producing the adaptations, of which spandrels are incidental by-products. It is responsible for producing structural changes in exaptations in order to fulfil their new functions. And it is responsible for maintaining exaptations in the population over evolutionary time even in the rare cases where no structural changes occurred.[33]

To deny that traits which are non-adaptive by-products have evolved would therefore also commit one to denying that human navels, or the whiteness of bone, are the products of evolution.

> There is no evidence that the belly button, *per se*, helped human ancestors to survive or reproduce. A belly button is not good for catching food, detecting predators, avoiding snakes, locating good habitats, or choosing mates. It does not seem to be involved directly or indirectly in the solution to an adaptive problem. Rather, the belly button is a by-product of something that is an adaptation, namely, the umbilical cord that formerly provided the food supply to the growing fetus.[34]

> [B]ones are adaptations, but the fact that they are white is an incidental by-product. Bones were selected to include calcium because it conferred hardness and rigidity to the structure (and was dietarily available), and it simply happens that alkaline metals appear white in many compounds, including the insoluble calcium salts that are a constituent of bone.[35]

---

[30] Gould and Lewontin (1979).
[31] Gould's usage of this terminology is inconsistent, which has created some confusion. Here, I follow the most widely accepted interpretations of these terms. See Buss et al (1998) for discussion.
[32] Gould (1991) p. 57.
[33] Buss et al (1998) p. 543.
[34] Buss et al (1998) p. 537.
[35] Tooby and Cosmides (1992) p. 63.

To deny that these traits have evolved is deeply counterintuitive, and introduces arbitrariness into the distinction between traits which have evolved and those which have not.

To summarise, traits which are non-adaptive by-products are just as much the result of natural selection as those which serve a clearly identifiable adaptive purpose. Even if Prinz's claim that morality is such a by-product is accepted, therefore, he is not entitled to make the further claim that morality did not evolve.

(iii) Morality is Not the Product of a Developmental System

As will be recalled from the previous chapter, accepting developmental systems theory means reconceiving evolution so that it includes any extra-genetic factors relevant to development. Nevertheless, this does not mean that all of an organism's traits must necessarily be said to have evolved. Organisms may possess numerous "individual traits", i.e. traits which are acquired as a result of chance events unique to the life-history of a particular organism.[36] For morality to be identified as non-evolutionary from a developmental systems perspective, it must be shown to be an individual trait.

Defending the claim that morality is an individual trait is highly unappealing, however. One alarming consequence of doing so would be that the widespread agreement on moral issues found within (and between) cultures would become all but inexplicable. This is because each individual must be thought of as arriving at all of their moral beliefs through a series of unpredictable, chance occurrences. To adapt the example given by Paul Griffiths and Russell Gray, moral beliefs would be like scars on one's thumb. Accounting for the substantial amount of agreement amongst moral beliefs actually found would, therefore, be like trying to account for how an equally substantial percentage of the population all came accidentally to acquire identically scarred thumbs. Without appealing to a single unifying explanation (which would count as part of a developmental system), any such attempt would be extremely implausible.

Likening moral beliefs to individual traits also undermines one of the most typical motivations for rejecting evolutionary accounts of morality, specifically, the desire to emphasise the importance of the role played by culture in human development. It will be recalled from the previous chapter that, as argued by Griffiths and Gray, "[m]any species-typical features of human psychology may depend critically on stably replicated features of human culture".[37] From a developmental systems perspective, therefore, eschewing evolutionary explanations also means denying the cultural transmission of complex behavioural traits. Clearly, this prospect is philosophically unappealing.

(iv) Summary

---

[36] See Chapter One §6.ii.
[37] Griffiths and Gray (1994) p. 302.

This section has discussed three different strategies for rejecting the claim that morality is the product of evolution. Each of these strategies has, for different reasons, been shown to be unsuccessful.

The first of these strategies is to claim that morality has not evolved because non-genetic factors contribute more to its development than do genetic ones. This approach fails because it rests on the mistaken belief that the causal contributions of genetic and non-genetic factors in development can be separated from one another, and then independently weighted. They cannot, and as a result the first strategy is not a viable option.

The second strategy attempts to reject evolutionary interpretations of morality by denying that morality is an adaptation. Whether or not a trait is an adaptation is irrelevant to the evolutionary status of that trait, however. Selection does not only produce adaptations, but a host of other by-products, some useful, others not. Nevertheless, all these by-products are said to have evolved, no less than the adaptations directly targeted by selection.

Finally, any attempt to deny that morality is the product of a developmental system should be found highly implausible. This is because the notion of a developmental system includes both cultural and genetic influences on development. Any aetiology of moral beliefs which excludes appeal both to genetic and cultural factors in their development is reduced to the claim that we each acquire all of our moral beliefs by accident. This is such a counterintuitive claim that it is unlikely ever to be seriously defended.

According to the preceding discussion, many putatively non-evolutionary theories of ethics, as evidenced by that of Prinz, are misconceived as such. They should in fact be understood as making a tacitly evolutionary claim. Focussing on a developmental perspective can help us to identify that misconception more easily. But does the ubiquity of evolutionary explanation mean that all ethical theories must ultimately be considered evolutionary? In fact, it does not. In the next section, I show why this is the case by proposing a better way of distinguishing between evolutionary and non-evolutionary ethical theories. Whilst my proposal does re-classify some moral theories, particularly in metaethics, as evolutionary, it avoids confusion by more clearly setting out how evolutionary theories of ethics differ from non-evolutionary ones.

4. Distinguishing between Evolutionary and Non-Evolutionary Ethics

The previous section argued that the notion of a developmental system, and therefore of evolution so defined, is ubiquitous among philosophically respectable aetiologies of moral belief. It does not follow from this ubiquity, however, that every ethical theory must be described as evolutionary.

For an ethical theory to be evolutionary, it must do more than simply assert that our moral beliefs are the product of evolution. Arguments about how those beliefs evolved, the selection pressures responsible for their evolution, and whether or not their possession should be understood

as an adaptation (modular or otherwise), are the purview of moral psychology. Research into these issues both can and should be of interest to the evolutionarily minded ethicist. Such work can, after all, reveal much about the nature of moral judgement which may be of philosophical importance.

However, moral philosophy is not solely concerned with the how and why of moral development: its primary focus, rather, is on normative and metaethical issues. Investigating the duties which we have to one another, and asking questions about the metaphysical nature of those duties, is what distinguishes ethics from moral psychology. An evolutionary aetiology of our moral psychology is no more an *ethical theory* than are any other developmental accounts to be found in the psychological literature.

This is a commonplace distinction, but it is one which the debate on evolutionary ethics has failed to emphasise. This failure can surely be linked to the huge amount of polemical heat generated by the politicised arguments both for and against the claim that our moral beliefs have evolved. During its nascent years, the term "evolutionary ethics" was simply applied wholesale to any theory which made this claim. This led to an unhelpful blurring of the distinction between psychological and philosophical perspectives on ethics. Re-emphasising that distinction allows us to see more clearly the difference between evolutionary moral psychology and evolutionary ethical theory. The latter must do more than claim that our moral beliefs have evolved. It must also show that particular normative or metaethical conclusions follow from accepting an evolutionary perspective.

Thus, an ethical theory should be considered evolutionary if it makes either of the following claims:

1) The fact that humans have evolved places us under specific moral obligations; or
2) Evolutionary data has specific implications for our conception of the nature and status of our moral beliefs.

Here, claim (1) pertains to normative ethics, and claim (2) to metaethics. I say more about each of these in §4.i below. In §4.ii I discuss what a genuinely non-evolutionary theory of ethics might look like, and argue that such theories are surprisingly few and far between.

(i) Normative and Metaethical Evolutionary Claims

Establishing that a normative conclusion follows from an empirical matter of fact is a risky philosophical enterprise, and ethical theorists are ill-advised to try to move too quickly from an empirical "is" to a moral "ought". Nevertheless, drawing normative implications from the fact of human evolution is an approach with well-known historical adherents, such as Herbert Spencer, and ought clearly to count as evolutionary. Obvious examples of claims of this sort would include statements such as "we have evolved to eat meat, therefore we ought not to be vegetarians", and "evolution promotes the survival of the fittest, therefore we should not give money to charities which try to alleviate starvation". There are, however, less obvious examples of normative evolutionary claims. It will be recalled that a conceptual expansion of the term "evolved" follows

from accepting developmental systems theory. This expansion means that *prima facie* non-evolutionary claims no longer count as such. For example, the claim that "according to the cultural norms my community subscribes to, I ought not to make lying promises; therefore I ought not to make lying promises" is also to be understood as evolutionary. This is because cultural norms are an important part of any developmental system, and constitute one of the means by which psychological traits are often acquired.[38]

Normative evolutionary claims like the ones given above are problematic. The fact that a behavioural trait has evolved is in no way a normative justification of that trait. If this were the case, species-typical dispositions towards out-group directed violence, or a sexual double standard, were they shown to have evolved, would be no less normatively endorsed than species-typical dispositions to care for one's children. It might be thought that some normatively relevant distinction between these behavioural traits could be made, perhaps by claiming that not all of our actions promote the good of the species. This attempt fails, however, because natural selection promotes behaviours which benefit individuals (or, more specifically, an individual's genes), rather than the species as a whole. To incorporate the idea of the "good of the species" into a normative schema is therefore to resort to a non-evolutionary *telos* as the standard of right action. In so doing, the position outlined by claim (1) has effectively been abandoned and the normative theory in question is no longer an evolutionary one.

The second way of incorporating evolutionary data into an ethical theory is to do so at the metaethical level. This approach holds far more promise than the first, and is the strategy typically employed by modern evolutionary ethicists. How might this be achieved? The following chapters provide detailed examples of evolutionary approaches to metaethics which have recently been endorsed. Rather than attempting to summarise the forthcoming discussions, therefore, I limit my remarks in this section to a schematic outline of what such an approach might look like.

The most readily apparent form of evolutionary metaethics is one which uses evolutionary data to defend a particular metaethical account of the nature of moral judgement. For example, evolutionary evidence suggests that moral judgements are likely to have evolved from our capacity to engage in reciprocal exchanges with conspecifics. If engaging in these exchanges was a consequence of increased selection for behaviour directed by pro-social emotions, then it could be argued that morality ought to be understood as a product of these pro-social emotions. Such an analysis could then be used in the metaethical debate surrounding cognitivist/non-cognitivist interpretations of moral discourse. If evolutionary data suggests that moral judgements are emotionally constituted, this information makes non-cognitivist metaethics more plausible. This argument has in fact been appealed to by Simon Blackburn in defence of quasi-realist expressivism.

> [I]f evolution selects for societies whose members have achieved some kind of co-ordination in their actions and feelings, [...] [then cognitivism] is relatively cumbersome. [...] [This is because] we must start by getting into shape to cognize values, and then we have to get into shape to make something of what we have cognized. How much

[38] In the terms of the previous chapter, this can be referred to as secondary value transmission. See Chapter One §4.i

simpler if there were just the underlying properties, and a properly conative sensitivity to them: a propensity to form attitudes and choices and policies in the light of them.[39]

A detailed discussion of this proposal would be out of place here. For present purposes, it suffices to note that evolutionary metaethics is an attractive alternative to normative evolutionary ethics. It does not encounter the objection of deriving facts from values, and the claim that evolutionary analyses of morality should have a bearing on our conception of moral discourse is compelling. As will be seen in later chapters, the metaethical applications of evolutionary data are not limited to the cognitivism/non-cognitivism debate, but extend to issues of moral realism and anti-realism, and to moral epistemology.

It should by now be apparent that, according to the view which I have advocated in this section, a non-evolutionary theory of ethics will be one which denies that the evolutionary process has any normative or metaethical significance. This seems, *prima facie*, to be a perfectly straightforward sort of claim. In fact, however, evolutionary data is often surprisingly hard to expunge, particularly from the field of metaethics.

(ii) Evolutionary Data is Difficult to Avoid

Of course, not all ethical theories are even tacitly evolutionary. This can be seen by considering a purpose-built example of a possible Divine Command Theory (hereafter DCT). The example has intentionally been formulated to draw on evolutionary data in its aetiological explanation of our moral beliefs. It is therefore consistent with the ubiquitous interactionism which, as argued in §2, is accepted by both evolutionary and (putatively) non-evolutionary theories alike. As will be seen, what makes DCT non-evolutionary is *not* a denial that such interactionism occurs; *nor* is it the claim that the genetic influence in that interaction is negligible. As shown in §3, these are confused responses, and do not constitute legitimate rebuttals of the evolutionary claim. Rather, DCT denies that the evolutionary processes responsible for our moral beliefs are of normative or metaethical significance. For that reason alone, it is non-evolutionary.

According to DCT, then, human beings have evolved. In addition to this, DCT claims that God has miraculously intervened at various stages throughout the course of human evolution. He has done so in order to guide our phylogenetic development in such a way as to enable us to be receptive to His Divine commands. Given this formulation of DCT, evolutionary processes clearly play an important part in any fully worked-out aetiology of our moral beliefs.

Yet DCT is by no means an evolutionary ethical theory. This is because evolution plays no role in the explanation either of the normative content of morality, or of the nature of our moral judgements. Here, normativity is generated entirely by the conformity of an action with one of God's commands: facts about evolutionary development are normatively irrelevant. Nor does the evolutionary narrative in DCT tell us anything about the nature of moral discourse. Has God guided human evolution in order to give us emotions that dispose us to act in accordance with His

[39] Blackburn (1995) p. 51.

commands, or has He done so to make us cognitively developed enough to recognise the moral force of those commands? Either of these views is compatible with DCT as I have sketched it. Furthermore, the evolutionary aspect of DCT's narrative does not have a bearing on debates over moral realism. The realist force of DCT's normative content is generated by the claim that God exists. But God's existence is not a fact about human evolution. A sceptical response to DCT would involve rejecting the existence of God, not disputing the theory's evolutionary narrative. Despite having plenty to say about the role of evolution in the development of our moral beliefs, then, DCT is not an evolutionary ethical theory.

It is fairly straightforward to see that many normative theories can un-controversially be deemed non-evolutionary. Given the argument of the previous section, according to which normative evolutionary ethical arguments are highly problematic, this is no bad thing. So, for example, it is possible to claim that an agent ought always to act so as to produce the greatest happiness for the greatest number. There is no evolutionary data at work in this claim: its normative content would not need to be revised should some new information about human evolution suddenly come to light.

It is much more difficult, however, to avoid drawing on evolutionary data at the metaethical level (although the example of DCT shows that this can in principle be done). Thus, if one wishes to claim that mind-independent moral properties exist, and that it is in virtue of these properties that our moral judgements are true or false, one must provide some account of how it is that we are able to recognise them. This explanation may be adaptive: recognising moral properties increases an organism's fitness. Alternatively, it may be non-adaptive: recognising moral properties just happens to be one of the things that creatures with our level of intelligence can do. Either way, the explanation will be an evolutionary one. Conversely, one may wish to deny that moral properties exist. Making such a denial, however, will necessitate providing some account of what moral discourse is discourse about. At this point, there are numerous ways in which evolutionary data may enter the narrative. For example, are moral judgements statements of subjective preference? If so, why do we have those preferences and not others? Fully answering this question will surely call for the use of evolutionary data.

The difficulty of avoiding evolutionary explanations at the metaethical level has not been sufficiently appreciated. Even Prinz, who purports wholly to reject evolutionary theories of ethics, tacitly draws on evolutionary explanations in his metaethics. He makes the claim that morality is emotionally constituted, and that our moral beliefs are culturally transmitted through practises of emotional conditioning. This conditioning:

> [m]ay allow us to construct behavioral norms from our innate stock of emotions. If caregivers punish their children for misdeeds, by physical threat or withdrawal of love, children will feel badly about doing those things in the future. Herein lie the seeds of remorse and guilt.[40]

Here, Prinz's claim that moral emotions are socially constructed from other, innate emotions places him squarely in the evolutionary camp. If the emotions out of which morality emerges are innate then, by Prinz's own lights, those emotions must have evolved.

---

[40] Prinz (2008b) p. 404.

(iii) Summary

Adopting a developmental systems perspective requires us to significantly increase the number of traits which we think of as being the products of evolution. As shown in §3, this expansion means that any reasonable aetiology of our moral beliefs will be evolutionary. This does not, however, mean that all ethical theories are necessarily evolutionary. In this section, I have argued that preserving the distinction between evolutionary and non-evolutionary theories of ethics can best be achieved by emphasising the different aims of evolutionary moral psychology and evolutionary ethics. Whilst plausible theories of ethics may be unable to avoid drawing upon evolutionary moral psychology (particularly in its expanded, developmental systems guise), they need not claim that those evolutionary processes are of any normative or metaethical significance. Theories which deny this significance should be deemed non-evolutionary. Whilst such a denial can be made without great difficulty at the normative level, evolutionary data is much harder to avoid when discussing metaethical issues. Accepting a developmental systems perspective *does* therefore mean taking a greater number of ethical theories to be evolutionary than was previously thought. This conceptual expansion includes putatively non-evolutionary theories such as Prinz's.

5. Babies and Bathwater...

This result of the previous section may well seem counterintuitive. Surely, it might be thought, evolutionary theories of ethics shouldn't be as hard to avoid offering as the narrative of this chapter suggests. If accepting developmental systems theory leads to this conclusion, then perhaps that is so much the worse for developmental systems theory. Is it not preferable to revert to something akin to standard interactionism,[41] rather than to accept that philosophers like Prinz are so mistaken about what it is that they are committed to?

Adopting a developmental systems perspective is, on this view, like throwing the baby out with the bathwater. Whilst standard interactionism may have its problems, perhaps we create an even bigger problem by abandoning it in favour of a theory which so inflates our conception of evolved traits.

In response to this objection, I argue that adopting a developmental systems perspective yields far greater gains than it incurs costs. It is, as argued, the best way to retain an interactionist perspective whilst avoiding the confusions discussed in §3. Furthermore, adopting a developmental perspective minimises the risk of making overly hasty assumptions about the extent to which complex behavioural traits such as aggression are developmentally canalised. There is always a necessary environmental input in the development of such tendencies. Identifying this input may be an important step towards mitigating its influence, and thereby perturbing the development of the

[41] See Chapter One §6.i

trait. Of course there is no guarantee that this will be possible, but developmental systems theory encourages us to avoid resignedly accepting that we are "naturally" selfish, aggressive, or the like.

Whilst developmental systems theory necessitates some conceptual revision of how we use the term "evolution", and of the traits which take to be the products of evolution, this revision is innocuous. Indeed, given the argument of this chapter (that endorsing standard interactionism can generate serious confusion over which traits have evolved and which traits have not) this conceptual revision is in fact virtuous. Conceding that morality is the product of evolution does not entail committing oneself to claims that might preferably be disavowed. It is important to see that both evolutionary moral psychology and evolutionary theories of ethics are compatible with a wide range of theoretical perspectives. Accepting the aetiological claim that morality is the product of evolution does not mean that one must deny or downplay the importance of cultural factors in moral development; nor does it commit one to the view that any normative consequences follow from an evolutionary interpretation of morality.

Although it is difficult entirely to dissociate metaethics from evolution, this association is also innocuous. To accept an evolutionary interpretation of morality is not to presuppose the truth of any particular metaethical perspective. The fact that evolutionary arguments both for and against moral realism[42] have been defended, as well as arguments for and against non-cognitivism,[43] is testament to this. What is of primary theoretical importance is not simply the claim that morality evolved; rather, it is the philosophical argument that relates this claim to a particular metaethical position. Describing an ethical theory as evolutionary does not, therefore, entail a tacit acceptance of a given philosophical perspective. With this in mind, the benefits of accepting a developmental systems approach to evolution can clearly be seen to outweigh the cost of the conceptual and terminological revision that goes with it.

6. Conclusion

In this chapter I have shown adherence to standard interactionism to be the source of confusion in the literature on the evolutionary status of morality. Adopting a developmental systems perspective allows us both to identify and to avoid this confusion. Thus, the claim that the environment is primarily causally responsible for the development of our moral beliefs rests on a misconception. The relative importance of genetic versus environmental factors in individual development cannot be specified, and even if this were possible, it would lead to arbitrariness in our distinction between traits which have evolved and those which have not. Neither can those traits which are not the adaptive products of targeted selection be said not to have evolved. Evolution produces by-products as well as adaptations, yet these are no less a result of the evolutionary process.

Evolutionary explanations are ubiquitous. This does not mean, however, that all ethical theories are *de facto* evolutionary. Re-emphasising the conceptual distinction between moral psychology and moral philosophy allows us to see that evolutionary ethical theories are importantly

[42] See, for examples, Rottschaefer and Martinsen (1990) and Street (2006) respectively.
[43] See, for examples, Blackburn (1995) and Rauscher (1997) respectively.

different from evolutionary aetiologies of our moral psychology. Unlike the latter, evolutionary theories of ethics endeavour to arrive at normative or metaethical conclusions based on an analysis of evolutionary data. The ubiquity of evolutionary explanations makes evolutionary metaethics hard (though not impossible) to avoid. Adopting a developmental perspective therefore entails reclassifying some putatively non-evolutionary approaches to ethics, such as that proposed by Prinz, as evolutionary. Although this reclassification may be somewhat counterintuitive, it avoids the generation of terminological confusion, and does not necessitate an endorsement of specific metaethical perspectives. It is therefore innocuous.

The argument of this chapter has shown that evolutionary ethics, properly conceived, should not be thought of as a fringe theory. Evolution cannot simply be discounted as an influence on the formation of our moral beliefs, and it is clear that this influence may be of important metaethical significance. What is considerably *less* clear, however, is the precise nature of this significance. How should we incorporate evolutionary data into our philosophical conception of the nature of moral judgement, and of the potential for that judgement to issue in true or false moral beliefs? This is the question which will be explored throughout the remainder of this thesis. The next chapter introduces Richard Joyce's evolutionarily-motivated argument for an error theory. If Joyce's sceptical challenge is a success, it will show that knowledge of morality's evolutionary origins not only undermines our confidence in our individual moral beliefs, but also in the notion of morality itself.

Chapter Three

A Sceptical Challenge: The Evolutionary Argument for Moral Error Theory

1. Introduction

This chapter introduces Richard Joyce's evolutionarily motivated defence of moral error theory. §2 sets the scene by drawing some conceptual connections between Joyce's argument and its historical predecessor found in the work of John Mackie. It also shows how both theories use evolutionary data to overcome the realist rebuttals motivated by arguments from epistemological conservatism.

Following this, §3 sets out the evolutionary aspect of Joyce's argument. It discusses the evolutionary selection pressures which he takes to underlie the evolution of pro-sociality, and introduces the distinction between pro-social emotions and genuinely moral judgements. §4 then discusses Joyce's account of moral judgement, and his conception of ordinary moral discourse and its pre-theoretical, non-negotiable commitments. According to this conception, the existence of categorical moral reasons is an in-eliminable aspect of ordinary moral discourse.

With these details in place, §5 goes on to outline Joyce's argument for an error theory. This argument is made by conjoining his evolutionary narrative with his conception of ordinary moral discourse. As will be seen, the result is Joyce's claim that we have no independent reason to believe in the existence of moral facts, and that their putatively categorical force cannot be accommodated within a naturalistic framework. We therefore have good reason to doubt that moral facts exist. Given that it is committed to the existence of such facts, however, moral discourse is fundamentally in error.

Having set out this argument, §6 considers some preliminary replies to Joyce. These ultimately prove unsuccessful, but the position endorsed by James Dreier in §6.iii suggests a way of responding to Joyce's claim that a naturalistic approach is unable to give a satisfactory account of the categorical force of moral judgements. In §7 this response is more fully developed. Joyce's scepticism concerning the existence of moral facts remains, but with a naturalistic account of the categorical force of moral judgement in place, the prospects of defending moral realism against that scepticism are significantly improved. This will be the task of each of the subsequent chapters.

2. Error Theory and Epistemic Conservatism

To endorse an error theory about a particular discourse, whether that discourse concerns morality, religion, or Martians, is to claim that a specific feature of that discourse is based on a mistake. Thus, the argument for an error theory is advanced by making two claims. The first of these claims attempts to establish that the target discourse is committed to a particular ontological or putatively factual proposition: "it is the case that P", for example, or "there is an *x*". The second claim purports

to show that this commitment is in some way problematic: "it is highly unlikely that P", "there is no evidence for the existence of *x*". If both these claims can successfully be made then, taken together, they will allow their exponent to claim that the target discourse is in error to the extent that it is committed to the proposition being disputed.

Error theories are not the exclusive domain of moral philosophers. Indeed, the most familiar forms of error theory are typically found outside of moral philosophy. For example, atheists are error theorists with respect to theistic forms of religious discourse: they claim that such discourse is committed to the existence of God, then go on to deny that there is good reason to believe in God's existence.

One of the most famous arguments for a moral error theory was made by John Mackie. Mackie argued that:

> ordinary moral judgements include a claim to objectivity, an assumption that there are objective values [...]. But [...] although most people in making moral judgements implicitly claim, among other things, to be pointing to something objectively prescriptive, these claims are all false.[1]

Mackie's denial of the objectivity of morality was motivated by two arguments: the argument from relativity, and the argument from queerness. According to the former, variation between the moral codes endorsed by different societies cannot be explained quasi-scientifically, i.e. as the result of disagreement concerning empirically discoverable objective moral facts. Rather, different moral codes "reflect people's adherence to and participation in different ways of life".[2] That is, societies approve of vegetarianism, say, because the members of that society are themselves vegetarian.[3]

Mackie's second argument, the argument from queerness, holds that objectively prescriptive values are not the sort of things which are likely to be found to exist. This is because, were there any moral facts, "they would be entities or qualities or relations of a very strange sort, utterly different from anything else in the universe".[4] Knowledge of moral facts, unlike knowledge of other sorts of fact, would, according to Mackie, be intrinsically motivational.

> An objective good would be sought by anyone who was acquainted with it, not because of any contingent fact that this person, or every person, is so constituted that he desires this end, but just because the end has to-be-pursuedness somehow built into it. Similarly, if there were objective principles of right and wrong, any wrong (possible) course of action would have not-to-be-doneness somehow built into it.[5]

Such intrinsically motivating principles, Mackie argues, are incredibly difficult to cash out in naturalistic terms. It is, he argues, "much simpler and more comprehensible [...] [to] replace the

---

[1] Mackie (1977) p. 35.
[2] Mackie (1977) p. 36.
[3] It is possible to imagine a purely historical explanation for this vegetarianism, which does not presuppose any normative motivation to be at work.
[4] Mackie (1977) p. 38.
[5] Mackie (1977) p. 40.

moral quality with some sort of subjective response".[6] Thus, according to Mackie, there is good reason to doubt the existence of objective moral facts. To the extent that moral discourse is pre-theoretically committed to the existence of such facts, therefore, it is committed to an error.

A detailed discussion of Mackie's arguments lies outside the scope of this chapter. Nevertheless, Mackie's position is worth noting, for two reasons. The first of these is one of historical influence: Joyce's argument for an error theory has important conceptual connections with Mackie's; in particular with Mackie's conception of ordinary moral discourse and its vulnerability to the argument from queerness. The second of these reasons, however, is that Mackie's account of the origins of morality *also* includes an evolutionary component. For Mackie, morality developed in order to make cooperative endeavours more palatable to each of the agents involved. Without it, "limited resources and limited sympathies together generate [...] conflict and an absence of what would be mutually beneficial cooperation".[7] What is needed to promote cooperation is the acquisition of pro-social sentiments.

> [T]he ordinary evolutionary pressures, the differential survival of groups in which such sentiments are stronger, either as inherited psychological tendencies or as socially transmitted traditions, will help to explain why such sentiments become strong and widespread.[8]

As will be seen, this is less developed than the evolutionary narrative offered by Joyce, but its presence raises an interesting question: Why do both Joyce and Mackie feel the need to supplement their argument for moral error theory with evolutionary data? Why are arguments from queerness (or the like) not sufficient to make morality's claims seem deeply problematic?

The answer to this question is that an evolutionary narrative enables arguments for moral error theory to overcome psychological dispositions towards epistemological conservatism. Briefly put, epistemological conservatism is the view that "an agent is in some measure justified in maintaining a belief simply in virtue of the fact that the agent has that belief".[9] This particular formulation is often thought to be too strong: it appears to offer a justification for retaining any extant belief whatsoever, even if there is absolutely no evidence to support that belief. Weaker formulations of epistemological conservatism have therefore been offered. Notable among these is the claim that an agent may be justified in continuing to believe a proposition even if she has forgotten the original grounds for the formation of that belief. This is the case if she knows that her belief formation processes are generally reliable. Given this knowledge, the agent is able to infer that her extent belief is likely to have been formed by that generally reliable process. She can therefore be said to have a "reason to continue believing [those] things [she already believes] simply because [she] happen[s] to believe them already".[10]

I will not now assess the merits of arguments surrounding the principle of epistemological conservatism. For the purposes of the present discussion, it is sufficient to note that Joyce takes responding to epistemological conservatism to be one of the tasks faced by the moral sceptic. The

---

[6] Mackie (1977) p. 41.
[7] Mackie (1977) p. 111.
[8] Mackie (1977) p. 113.
[9] Christensen (1994) p. 69.
[10] Christensen (1994) p. 76.

evolutionary aetiology of morality which he provides is designed to fulfil this task. Thus, according to Joyce:

> a genealogical explanation serves to defeat or block whatever *prima facie* justification these [in this case moral] intuitions might otherwise have been granted. [...] [These intuitions are defeated] if one has a plausible, or even empirically confirmed, theory of where the intuitions in question come from that is consistent with their being false.[11]

The theory which Joyce uses as a means to this end, then, is the theory of evolution. Consequently, before giving a more detailed account of Joyce's argument for an error theory, it will be appropriate to set out the evolutionary aetiology of pro-social emotions designed to prepare us for that argument.

3. The Evolutionary Origins of Pro-Social Emotions

Explaining why an organism expends time and effort in attracting a mate, or in foraging for food, poses no difficulty for Darwinism: such behaviour clearly increases the genetic fitness of the agent, and natural selection favours organisms that act so as to increase their fitness. Dispositions to act in fitness increasing ways are precisely the sort of thing which evolution can be expected to produce.

Yet organisms also behave in ways which seem to *reduce* their genetic fitness. For example, numerous species of birds, mammals, and primates have evolved to give alarm calls. These calls warn nearby conspecifics of the presence of a predator. This has the effect not only of providing an early-warning for those conspecifics, but also risks drawing attention to the individual giving the alarm call, thereby increasing the likelihood of its being predated. Alarm-calling does not, therefore, appear to be in the genetic interest of the agent making the call: why not simply flee, leaving one's less observant conspecifics to face the danger? Equally puzzling, from a genetic perspective, is the practice of blood-meal sharing found in vampire bat colonies. If these bats do not feed regularly, they starve. Yet the success of any given night's hunt is strongly influenced by chance factors. Bats unable to find food will return to the colony and solicit blood from conspecifics that have had better luck. These successful hunters will regurgitate some of the blood which they have consumed, effectively providing a free meal for their hungry comrades.

Behaviour which increases the fitness of a second party, whilst decreasing the fitness of the agent, is described as "genetically altruistic". Various explanations for the persistence of putatively genetically altruistic behaviours have been suggested, and Joyce argues that the evolution of pro-social emotions is closely connected to the evolution of altruistic behaviour. His discussion of the latter draws on five explanations which have been offered for the natural selection of altruistic behaviour, and I consider each of these in turn.

(i) Kin Selection

[11] Joyce (2010) p. 45. Emphasis added.

Kin selection theory was first proposed by William Hamilton. Hamilton did not dispute that natural selection favours behaviour which increases the likelihood of an organism's genes being transmitted to the next generation. Yet he observed that an agent's genes were also capable of being transmitted by organisms other than the agent herself. This is because those same genes are likely to be found in conspecifics to which the agent is closely related.

> [F]or a gene to receive positive selection it is not necessarily enough that it should increase the fitness of its bearer above the average if this tends to be done at the heavy expense of related individuals, because relatives, on account of their common ancestry, tend to carry replicas of the same gene; [...] a gene may receive positive selection even though disadvantageous to its bearers if it causes them to confer sufficiently large advantages on relatives.[12]

Hamilton introduced a distinction between classical fitness and inclusive fitness. The former is the familiar measure of an organism's reproductive success. If classical fitness were all that was selected for, behaviours such as alarm calling would be an evolutionary enigma. But this is not the case. Inclusive fitness is also selected for.

> Inclusive fitness is calculated from an individual's own reproductive success plus his effects on the reproductive success of his relatives, each one weighted by the appropriate coefficient of relatedness.[13]

Thus, if giving an alarm call is likely to allow numerous relatives to escape from potential harm, it may be in the genetic interest of the potential alarm caller to do so, even if this is at the cost of her own life.

Joyce notes that human pro-sociality is by no means restricted to close kin: indeed, "the tendency to favour one's own family members is a vice to which we have given the name 'nepotism'".[14] Nevertheless, he argues that kin selection may have played an important role in the evolution of pro-social behaviour aimed at non-kin. This is because the basic psychological capacity for pro-sociality generated by kin selection may have constituted the evolutionary foundations of less discriminate forms of pro-social behaviour.

> Biological natural selection is a conservative process, bending old structures into new, pressing into service available material for novel purposes. [...] If kin selection gave our distant ancestors the psychological and physiological structures needed for regulating helpful behavior toward family members, then those structures became available for use in new tasks – most obviously, helpful behavior toward individuals outside one's family – if the pressures of natural selection pushed in that direction.[15]

---

[12] Hamilton (1964) p. 17.
[13] Dawkins (1982) p. 186.
[14] Joyce (2006) p. 21.
[15] Joyce (2006) p. 22.

(ii) Mutualism

One of the selection pressures for expanding one's repertoire of pro-social behaviours is the increase in fitness that can be gained by engaging in mutualism. Joyce's discussion of mutualism is brief. His remarks are largely confined to the observation that certain sorts of agential endeavour stand more chance of being successful when undertaken as part of a group. Thus:

> [a] lion might want a piece of elephant for dinner [...] but one lion will not be able to accomplish this by itself. If a group of lions find themselves in this situation, they will do well by cooperating in the bringing down of an elephant. If they don't cooperate, all of them will go hungry [...].[16]

Joyce does note, in addition to this, that engaging in mutualism does not presuppose that repeated interactions occur between individual organisms. This means that mutualism is conceptually distinct from reciprocal altruism, the third item in Joyce's discussion.

(iii) Reciprocal Altruism

Reciprocal altruism was first discussed by Robert Trivers.[17] Trivers argued that an agent could sometimes increase her fitness in the long term, by making a sacrifice for a conspecific which reduces her fitness in the short term. This will be the case when the agent can expect the conspecific on whose behalf she has made that sacrifice to make a similar sacrifice on *her* behalf at a later date. This somewhat abstract characterisation can be made more lucid by returning to the example of blood-meal sharing in vampire bats, discussed above.

Imagine two bats, Donor and Recipient. One night, Donor manages to find food, whereas Recipient does not. Recipient successfully solicits blood from Donor, whose fitness is somewhat reduced by the sacrifice of part of her food supply. Note, however, that Recipient's fitness is increased by a *greater* amount than Donor's is reduced. This is because Donor does not sacrifice *all* of her food resources, and so does not expose herself to the risk of starvation. Recipient, on the other hand, has no food sources to begin with, and so is *already* at risk of starvation. Donor's sacrifice is thus worth more to Recipient *at that time* than it is to Donor. Several nights later, however, their roles are reversed. It is now Recipient who has been successful in the hunt, and Donor who is at risk of starving. Donor therefore solicits blood form Recipient, who, having received a supply of blood from Donor in the past, reciprocates her previous sacrifice. Because of the role reversal, the blood which Donor now receives is worth more to her than the blood which she had previously donated. Both parties, therefore, gain from engaging in reciprocally altruistic exchanges.

At this point, however, a complication arises. Why does Recipient return the blood-sharing favour? Why not simply take blood from Donor, and then opt-out of the reciprocal relationship? This

---

[16] Joyce (2006) p. 22.
[17] See Trivers (1971).

would confer the greatest increase in fitness, as Recipient has accepted a benefit without having to pay the associated cost. The answer to this question lies in the open-ended iteration of the interactions between Donor and Recipient. If Recipient is unlikely to have anything to do with Donor in the future, then it makes adaptive sense for Recipient to refuse to reciprocate. But life in a vampire bat colony is not like this. Because successful hunting is heavily influenced by luck, the chances are good that, in the fullness of time, Recipient will need to solicit food once again. But if Recipient has previously spurned Donor's solicitations for food, Donor will refuse to give up any of her food supply to Recipient. At this point, Recipient is in serious risk of starvation. It is therefore in the interest of each of the vampire bats in the colony to reciprocate when called upon to do so. The increase in fitness that comes from so doing outweighs the costs associated both with not engaging in blood-sharing at all, and also with engaging but refusing to reciprocate.[18]

Joyce, following Richard Alexander, refers to this form of reciprocal altruism as "direct reciprocity", and distinguishes it from "indirect reciprocity". Indirect reciprocity is different from the direct variant in that the benefit received by the agent who makes the initial sacrifice (i.e. Donor, in the above example) is not conferred by the original recipient of that sacrifice.

Indirect reciprocity works by the cultivation of reputation. Essentially, the idea is that by being seen to make some costly sacrifice, an agent can be identified as someone with whom it is desirable to engage in a reciprocally advantageous relationship. As Alexander explains:

> [i]ndirect reciprocity involves reputation and status, and results in everyone in a social group continually being assessed and reassessed by interactants, past and potential, on the basis of their interactions with others.[19]

Joyce claims that indirect reciprocity is "of central importance"[20] to an account of the evolutionary origins of morality. This is because practices of indirect reciprocity can make it vitally important for an agent to establish a good reputation for herself: not only does having a good reputation make it more likely that she will engage in rewarding social exchanges, it also reduces her chances of becoming socially ostracised as a non-cooperative member of the group. Indirect reciprocity therefore creates novel selection pressures, which in turn "can lead to the development of just about any trait – extremely costly indiscriminate helpfulness included".[21]

(iv) Group Selection

The fourth strand in Joyce's evolutionary aetiology of morality is group selection. "Group selection" is a term with potentially misleading connotations. Historically, the term was used to describe behaviours performed solely for the good of the group, that produced no benefit whatsoever for the agent. This is *not* how the term is now used, however. Such self-sacrifice cannot

---

[18] For a more detailed discussion, see Dawkins (1976) Chapter 12.
[19] Alexander (1987) p. 85.
[20] Joyce (2006) p. 31.
[21] Joyce (2006) p. 33.

evolve, as it reduces the fitness of the sacrificing agent in both the short- and long-term, with no compensatory payoff (unlike the behaviours in the examples discussed above).

Confusingly, however, there are multiple definitions of group selection currently being used in the literature. For example, group selection has been used to refer to the fitness gains which accrue to each member of any set of interacting individuals engaged in a cooperative partnership, however transitory that partnership may be. According to such a definition, kin selection, reciprocal altruism, and mutualism are all forms of group selection. This is the view presented by Elliott Sober and David Sloan Wilson.[22] Samir Okasha describes this approach as "controversial", however, and notes that "[their] argument is conceptual rather than empirical; it turns on the meaning of 'group selection' and how it should be modelled".[23] In addition to this definition, group selection has also been said to be operational in what is ordinarily thought of as individual selection. This is because "an organism is itself a group of cooperating cells, and a eukaryotic cell is a group containing organelles and the nuclear chromosomes".[24] Joyce has neither of these definitions in mind during his discussion of group selection, however, and so they may be set aside without further discussion.

When Joyce talks of group selection, he means to describe the differential selection of groups containing a percentage of helpers in their population. For this sort of group selection to take place, very specific environmental conditions and population dynamics must obtain. Firstly, it is assumed the co-operators in a group will act so as to increase the overall fitness of the group of which they are a part by a constant and indiscriminate rate. That is, their actions will benefit every member of the group, and will not be directed solely at reciprocating individuals, as in the previous examples. Secondly, it is supposed that co-operators can identify one another, and will preferentially seek out the company of other co-operators, rather than non-cooperative members of the group. This preference generates a tendency for the group gradually to divide into two smaller sub-groups. One of these sub-groups will, for the most part, be composed of cooperative individuals who have preferentially associated with one another; the second will therefore mostly comprise non-cooperative individuals. Both of these sub-groups, it is then stipulated, suffer equally proportional losses in their populations. That is, two-thirds (say) of both sub-groups die, with individuals from both sub-groups facing an equal chance of numbering among the casualties. This could be for any number of reasons, such as predation, famine, or disease. Owing to their reduced sizes, the two sub-groups then re-merge. As a result, the overall percentage of co-operators in the newly merged group will be greater than the percentage of co-operators in the original group, prior to its initial division. If this process is repeated enough times, the final merging will result in a population which consists entirely of co-operators. Furthermore, because co-operators have a tendency to associate with one another, a homogenous population of co-operators will no longer exhibit a tendency to divide into sub-groups.

As is apparent even from this short summary, the conditions in which such group selection is likely to take place are highly specific. Indeed, Joyce notes that the "complain[t] that this is all just fiddling with numbers to get the desired result [...] [contains] a hint of truth".[25] Nevertheless, he

---

[22] In Sober and Wilson (1998).
[23] Okasha (2006) p. 177.
[24] Okasha (2006) p. 173.
[25] Joyce (2006) p. 37.

argues, group selection illustrates one of the ways in which cooperation "*could* develop through the forces of biological natural selection".[26]

The attraction of (this version of) group selection theory is that it purports to show how a population of indiscriminate co-operators could be naturally selected for. It should be noted, however, that the theory *presupposes* that the original group *already* contains at least some indiscriminate co-operators. Whilst group selection theory may, therefore, show how such ultra-cooperative individuals might gradually come to supplant the non-cooperative members of their group, it does *not* account for the genesis of indiscriminate cooperation. It is therefore likely that some form of kin selection, mutualism, or reciprocal altruism will need to be appealed to in order to account for the existence of the cooperative behaviour which group selection selects for. Indeed, Joyce himself is ambivalent about the role which group selection is likely to have played in the evolution of morality: he ends his discussion with the codicil that "it is questionable how large a role [group selection] played in human ancestry".[27]

(v) Cultural Group Selection

Joyce proposes the four types of selection pressure discussed above as possible candidates for an explanation of the evolution of basic forms of cooperation. Yet he does not take them to be sufficient explanations for the highly developed forms of cooperation found in human societies. Following his discussion of group selection, then, he proposes that *cultural* group selection may be a better explanation of the existence of specifically human forms of cooperation. As cultural selection has already been discussed in some detail in chapter one, I will only briefly outline Joyce's comments in this section.

Joyce notes that a conformist bias in a cultural group would lead each of the members of that group to adopt similar social norms. This process would generate selection pressure between cultural groups whose norms significantly diverged, in turn allowing groups whose norms included the advocacy of peaceful cooperation to be preferentially selected over groups without this norm.

> Once there exists a meta-population of culturally distinct groups, there is selective pressure in favor of the persistence and proliferation of those cultural traits that are broadly "prosocial".[28]

(vi) Helping Behaviour and Psychological Altruism

---

[26] Joyce (2006) p. 37. Original emphasis.
[27] Joyce (2006) p. 41.
[28] Joyce (2006) p. 43.

With the exception of group selection, whose efficacy can anyway be disputed, the evolutionary mechanisms discussed above do not so much seem to *explain* genetic altruism, as to explain it *away*. Alarm calling, blood-meal sharing, and other forms of cooperative endeavour are seen to be in the genetic interests of the agents who perform them, rather than the selfless acts of sacrifice which they *prima facie* seem to be. This is something of which Joyce is well aware. He terms the actions produced by these mechanisms as instances of "helping", and his definition of helping as "[b]ehaving in a way that benefits another individual"[29] pointedly makes no mention of a reduction in the agent's genetic fitness.

There is, however, no necessary connection between genetic altruism and *psychological* altruism. Accordingly, Joyce argues that genetically selfish actions may nevertheless be prompted by psychologically altruistic emotions.

> If human reproductive fitness was enhanced by a proclivity for helping family members [...], what might the process of natural selection have done to our brains in order to accomplish this? An important part of the answer, I think, is clear, simple, and rather agreeable: love.[30]

The first chapter of this thesis[31] raised the question of whether an evolutionary perspective on human behaviour entails that all of our actions are ultimately self-interested. We are now in a position to see why it does not.

> The possibility that love (say, a father's love for his child) might be given an evolutionary explanation of the kind just provided does not imply that the father's love is "really selfish" on the grounds that it is motivated by an unconscious desire to optimize his own inclusive fitness. [...] To be ignorant of an element of the explanation of why you are feeling love – to be ignorant of the evolutionary ancestry of the emotion, for example – is not to be mistaken about the true object of your emotion [...].

Joyce has thus set out an evolutionary aetiology of genuinely pro-social emotions, derived from his discussion of the evolution of helping behaviour. It is worth noting at this point that Joyce's aetiology is not a controversial one. Most discussions of the evolution of morality argue that it has its evolutionary origins in selective processes such as kin selection and reciprocal altruism. Furthermore, even if evidence should later come to light which calls this account into question, it will not radically undermine the strength of Joyce's argument. It will be recalled that Joyce's motivation for providing an evolutionary aetiology is to overcome psychological tendencies towards epistemic conservatism. To that end, as Guy Kahane observes, "it does not matter here whether any *particular* evolutionary explanation is true. What matters is that *some* such story is likely to be true".[32]

Nevertheless, Joyce does not take it that he has given an evolutionary aetiology of *morality*. This is because there is much more to morality, on his account, than the possession of a disposition to act pro-socially: "someone who acts solely from the motive of love or [some other form of

---

[29] Joyce (2006) p. 13.
[30] Joyce (2006) p. 47.
[31] See §2
[32] Kahane (2011) p. 111.

psychological] altruism *does not thereby make a moral judgment*".[33] To see why this is the case, it is necessary to consider Joyce's account of moral discourse. This, then, will be the task of the next section.

4. Joyce's Conception of Moral Discourse

(i) The Dual Function of Moral Judgements

Joyce by no means takes the aetiology of pro-social emotions outlined above to be irrelevant to an understanding of morality. Indeed, he claims that "social emotions [...] are of central importance to human morality".[34] Despite their importance, however, pro-social emotions are not sufficient for the existence of genuine morality. This is because to be moral, for Joyce, requires more than acting out of a pro-social disposition; it also requires an awareness of prohibitions: "the idea that one shouldn't kill or steal because to do so is wrong".[35] Pro-social emotions can equip us with inhibitions against killing and stealing, but moral *prohibitions* are importantly different from these. Unlike non-cognitive inhibitions, prohibitions contain a cognitive component.

Yet, according to Joyce, moral judgements are no more wholly cognitive than they are wholly non-cognitive. That is, he argues that "moral judgments (as speech acts) express both beliefs and conative non-belief states".[36]Joyce draws support for this claim via a discussion of Moore's paradox. Moore's paradox describes a feature of statements comprising non-contradictory sentence pairs, where the effect of the second sentence is to nullify the content of the first. An example of such a statement would be "the cat is on the mat; but I don't believe it".[37]

Joyce argues that moral judgements can be framed in ways that exemplify Moore's paradox, and that by attending to each of the sentence pairs comprising such judgements, morality can be seen to have specifically cognitive features. Thus, Joyce argues that if someone were to say "Hitler was despicably evil. But I don't believe that he was despicably evil",[38] it would be reasonable to interpret the second sentence as a denial of the first. It follows from this interpretation, Joyce argues, that the first sentence functions as "an expression of belief – that is, an assertion".[39] Nevertheless, Joyce argues, moral judgements are not wholly cognitive. Pure cognitivism, according to Joyce, is "clearly inadequate".[40] This is because moral judgements are typically used, at least in part, to convey information about the speaker's attitude towards the agent or action which is the subject of the judgement.

---

[33] Joyce (2006) p. 50.
[34] Joyce (2006) p. 51.
[35] Joyce (2006) p. 50.
[36] Joyce (2006) p. 56.
[37] Joyce (2006) p. 55.
[38] Joyce (2006) p. 55.
[39] Joyce (2006) p. 55.
[40] Joyce (2006) p. 57.

> To say that Mary is *morally bad* is not like making a neutral statement, such as "She took the book from the shop without paying for it". In this latter statement one's attitude toward Mary is left uncommitted; it is possible that the speaker is an anarchist who heartily approves of shoplifting. But if the sentence "Mary is morally bad" is seriously put forward [...] we would know that the speaker disapproves of Mary.[41]

Thus, for Joyce, a moral judgement expresses both a belief *and* some attitude held by the speaker who utters it.

(ii) Non-Negotiable Features of Moral Discourse and the Linguistic Nature of Morality

The claim that moral judgements express beliefs is an essential part of Joyce's error theory. What they express beliefs *about*, he argues, are putative reasons for action. Joyce argues for this claim by discussing what he calls the "practical clout" of moral judgements.

That moral judgement ostensibly provides reasons for action can be seen from the way it is used to influence the behaviour of others. Suppose, for example, that an ardent campaigner against animal cruelty attempts to stop a fox-hunt. If she did so simply by proclaiming loudly that she did not like fox-hunting, or else "by yelling 'Boo to fox hunting!'",[42] she would not thereby have provided the hunters with a reason to refrain from hunting foxes.[43] Suppose, however, that she were to shout: "Fox hunting is morally wrong. Your actions are evil!"[44] In this case, the protester *has* provided the fox-hunters with a putative reason to refrain from such activity: specifically, its moral wrongness.

Joyce argues, in addition, that moral judgements are often used *categorically*. That is, they present reasons for action which apply to any agent, irrespective of whatever other desires or projects that agent may have. Thus, when hunters are faced with the objection that fox-hunting is morally wrong, to simply reply that they very much enjoy fox-hunting, and should therefore continue to do it, would be to miss the point of the objection. As Joyce puts it:

> moral imperatives are not merely the ones that people *do not* evade by citing special goals, they are the ones that people *cannot* evade by citing special ends. This is what makes them *inescapable*.[45]

Joyce emphasises that he does not mean here to refer to a specifically Kantian conception of categorical imperatives. Rather, his use of the term is simply to be understood as referring to "an imperative that does not recommend a means to an end".[46]

---

[41] Joyce (2006) p. 57. Original emphasis.
[42] Joyce (2006) p. 58.
[43] Assuming, that is, that the hunters were not previously disposed to avoid incurring the protestor's displeasure.
[44] Joyce (2006) p. 58.
[45] Joyce (2006) p. 61. Original emphasis.
[46] Joyce (2006) p. 61.

Joyce holds that both the assertoric nature of moral judgements and their capacity to be used categorically are in-eliminable components of moral discourse. That is, they are "non-negotiable"[47] features of moral discourse: without them, moral discourse would be unrecognisably altered. In earlier work,[48] Joyce provides a heuristic for conceptualising this non-negotiability. He does this by describing a "translation test". According to this test, one takes a candidate term from a particular language, and considers whether a proposed English language translation of that term (for example) picks out the same concept. When a proposed translation is found to be inadequate, Joyce argues, it is because the translation fails to convey one or more non-negotiable features of the concept which the term being translated denotes. Numerous terms have been abandoned throughout the course of history, Joyce argues, because a non-negotiable part of the concept which they denote has been found to be irredeemably problematic. For example, it is impossible, according to Joyce, that discourse about the theoretical phlogiston particle (once thought to be responsible for combustion) could have been translated into discourse about oxygen particles, once they were discovered.

> [T]he discovery that we had been wrong in thinking that there is a stuff stored in combustible bodies and released during burning was sufficient for us to decide that there is no phlogiston at all. When Lavoisier gave us the concept *oxygen*, it wasn't available for Stahl to say "Well this stuff that Lavoisier is calling 'oxygen' just *is* what I've been calling 'phlogiston' all along – I was just mistaken about its being stored and released during combustion." The belief that phlogiston is stored and released was a *non-negotiable* part of phlogiston discourse – the falsity of this belief was sufficient to sink the whole theory.[49]

Joyce holds that moral discourse is similarly committed to the assertoric and (at least potentially) categorical nature of moral judgement. Yet these are not the only non-negotiable features of moral discourse which he identifies. Joyce in fact identifies a total of nine such features, and cites these as evidence that "the linguistic conventions that govern moral discourse are those of assertions".[50] These nine features are:

(1) They (moral utterances) are expressed in the indicative mood.
(2) They can be transformed into interrogative sentences.
(3) They appear embedded in propositional attitude contexts.
(4) They are considered true or false, correct or mistaken.
(5) They are considered to have an impersonal, objective character.
(6) The putative moral predicates can be transformed into abstract singular terms (e.g., "goodness"), suggesting they are intended to pick out properties.
(7) They are subject to debate which bears all the hallmarks of factual disagreement.
(8) They appear in logically complex sentences (e.g., as the antecedents of conditionals).
(9) They appear as premises in arguments considered valid.[51]

---

[47] Joyce (2001) p. 3.
[48] Joyce (2001).
[49] Joyce (2001) p. 4.
[50] Joyce (2001) p. 13.
[51] Joyce (2001) p. 13.

Joyce sees moral discourse as actively constituted by these linguistic conventions. Any discourse which lacked one of them would not be a genuine *moral* discourse. This means that for Joyce, morality is an essentially linguistic phenomenon. Indeed, he states this explicitly when he claims that "language is a prerequisite for having moral concepts".[52]

It is not possible for an agent to make competent moral judgements unless that agent is a language user, Joyce holds. This, he argues, is because an essential aspect of making such judgements involves consciously subscribing to the norms which they are used to express.

> Just as one cannot be counted competent with the word "kraut" unless one knows that it is a derogatory word, one cannot be counted competent with moral language unless one knows that in using it (seriously) one expresses subscription to certain practical standards.[53]

In addition to this competency requirement, moreover, Joyce holds that certain emotions are made possible only through their being realised in a linguistic medium. One such emotion is disgust. Joyce claims that disgust is likely to have evolutionary origins in reflexive mechanisms designed quickly to expel harmful foodstuffs from the mouth, thereby preventing their consumption. Yet specifically moral disgust is importantly different from these mechanisms, and this difference is generated by an additional conceptual-linguistic component. Thus, disgust:

> includes a feeling of offensiveness and contamination. The latter is well documented by experiments showing subjects to be reluctant to wear a sweater that has been used by a stranger even after it has been scrupulously laundered – their willingness plummeting further if the owner is believed to have had an amputated leg, or to have committed murder [...].[54]

Joyce argues that such research shows disgust to have an important cognitive component, centred on the existence of invisible contagions. Without this conceptual supplementation, the basic evolutionary mechanisms that underlie disgust would not be capable of producing that emotion.

Importantly, certain moral emotions also depend upon conceptual supplementation for their existence. Thus, for example, the emotion of guilt is associated "with the judgment that the person has performed a wrongful action for which amends might be made";[55] whereas shame is produced by "a judgment that the person is 'wrong in himself', and perhaps nothing can be done about it".[56]

For Joyce, then, moral judgements are constituted by a synthesis of evolved emotional dispositions, coupled with a linguistic-conceptual component which fundamentally alters their nature. It is in virtue of this linguistic element that human morality is made possible: we are not simply disposed to perform certain sorts of actions, and to avoid performing others. Rather, we conceive our actions as subject to prescriptive regulations, as potentially violating norms to which we subscribe. Furthermore, these norms are inescapable, factual entities: our beliefs about them

---

[52] Joyce (2006) p. 76. Emphasis removed.
[53] Joyce (2006) p. 84.
[54] Joyce (2006) p. 96.
[55] Joyce (2006) p. 102.
[56] Joyce (2006) p. 102.

can be true or false, and they apply to us irrespective of our personal projects and desires. This assertoric, categorical nature of moral discourse is a non-negotiable feature of it: any discourse about norms that lacked these features would not be something that we could recognise as a genuinely *moral* discourse.

Having outlined Joyce's conception of moral discourse, it is now possible to show how he argues that, when conjoined with an awareness of morality's evolutionary origins, such an account leads to a moral error theory.

5. Joyce's Argument for Moral Error Theory

Joyce's argument for an error theory is based on an epistemological claim. He begins by arguing that: "[w]ere it not for a certain social ancestry affecting our biology [...] we wouldn't have concepts like *obligation, virtue, property, desert*, and *fairness* at all".[57] Our normative endorsement of these concepts is in part a result of the evolutionary processes which created the human mind. An awareness of this, Joyce argues, gives us good reason to doubt the objective prescriptiveness built into these concepts. But why should this be the case? Why should an awareness of morality's evolutionary origins call the validity of moral discourse into question?

Joyce's answer to this question is that the non-negotiable features belonging to moral discourse have in fact been instilled in our psychologies by natural selection: "the very concept of 'requirement' is the product of natural selection".[58]

> When, for example, I judge that slavery is forbidden, certainly the concept *slavery* is something that I have acquired, but the concept *forbidden* is not. It would be a mistake to conclude from this that my belief that slavery is wrong is the product of natural selection. Rather, it is my belief that there exist forbidden actions (or that forbiddenness exists) that is so produced.[59]

This poses a problem for the realist commitments of moral discourse, Joyce argues, because the fact that we have evolved to believe some actions to be forbidden (or required, or fair, etc.) is not a reliable indication that those actions *are* genuinely forbidden. "It was no background assumption of [the evolutionary aetiology of morality] that any actual moral rightness or wrongness existed in the ancestral environment."[60]

Joyce uses an analogy to develop this argument more fully. He asks his reader to consider the existence of a belief pill, consumption of which will inevitably induce the belief that Napoleon lost the battle of Waterloo. Joyce then asks whether, upon discovering "that at some point in your past someone slipped you a 'Napoleon lost waterloo' belief pill",[61] this should cause you to doubt

---

[57] Joyce (2006) p. 181.
[58] Joyce (2001) p. 162.
[59] Joyce (2001) p. 162.
[60] Joyce (2006) p. 183.
[61] Joyce (2006) p. 179.

the truth of that belief. For Joyce, the answer to this question is clear: "Of course it should".[62] This is because in such a scenario, the means by which you have come to acquire the belief that Napoleon lost Waterloo are unreliable. The fact that you have taken such a pill has no obvious relation to the events of 18th June 1815. Joyce is careful to note that *it may be the case* that Napoleon lost Waterloo. Taking a belief pill no more tells against this proposition than it does in favour of it. Still, in order to have any epistemic confidence in our belief about the outcome of the battle, we would need to uncover some independently confirmatory or dis-confirmatory evidence.

Joyce argues that an evolutionary aetiology of morality should inspire the same scepticism with regard to our moral beliefs as does the realisation that we have been slipped a belief pill with regard to our certainty about the outcome of Waterloo:

> once we become aware of this genealogy of morals we should [...] cultivate agnosticism regarding all positive beliefs involving these concepts until we find some solid evidence either for or against them.[63]

Indeed, according to Joyce this scepticism should extend not only to the content of our positive moral beliefs, but to our acceptance of the concept of morality itself.

> It is not a matter of allowing oneself to have an open mind about, say, the wrongness of abortion [...]; rather it is a matter of maintaining an open mind about whether there exists *anything* that is morally right and wrong [...].

This latter claim allows us to see where Joyce's analogy with the belief pill breaks down. Whilst realising that I have taken a belief pill should cause me to question whether Napoleon did in fact lose the battle of Waterloo, Joyce does not claim that it should cause me to question *whether that battle actually took place*. If the analogy to moral beliefs were exact, then an evolutionary aetiology should cause me to question whether abortion is morally right or wrong; yet it need not prompt me to question *whether there is such a thing as* moral rightness and wrongness. This latter position would amount to the claim that evolution has a biasing influence on the sort of moral judgements which we tend to form, and that this influence should be corrected for. Of course, such a position is much easier to respond to: one could simply argue that detecting and correcting for just such a bias (albeit not conceived as an evolutionary one) has been the task of moral philosophy for millennia. Yet this is not Joyce's claim. To make the belief pill analogy exact, it should be specified that the pill not only induces a belief about the outcome of the battle; it also induces the belief that the battle took place. Again, this is not say that the battle did not occur, or that its outcome was otherwise than as specified by the belief pill. It is possible to imagine a society comprising individuals, some of whom have unwittingly taken a belief pill, and others of whom have not, yet each with identical true beliefs concerning the battle and its outcome. Still, knowledge that one belongs to the set of individuals who have taken a belief pill should cause one to suspend acceptance of all of one's beliefs about the battle of Waterloo, including its actual occurrence. Once this amendment to the analogy had been made, Joyce's claim can be seen to apply to the existence of moral truths *tout court*, and not merely to the putative factuality of positive moral claims.

---

[62] Joyce (2006) p. 179.
[63] Joyce (2006) p. 181.

How, then, does Joyce proceed from this epistemological argument to a moral error theory? According to his argument, the evolutionary aetiology which he has provided allows non-natural and supernatural accounts of moral judgement simply to be discounted. This is because "non-naturalism and supernaturalism [...] posit extra ontology in the world, but the presence of the non-moral genealogy [...] shows this ontology to be explanatorily superfluous".[64] That is, positing a realm of non-natural or supernatural moral facts does not contribute anything to a genealogy of moral belief that cannot also be contributed by a strictly naturalistic approach. Thus, "Ockham's Razor [...] can come in and do its thing".[65] This leaves naturalism as the only perspective from which to offer a realist account of moral judgment. Yet Joyce argues that any such attempt is likely to fail. In doing so, he supplements his epistemological argument with an appeal to a Mackie-esque argument from queerness. Joyce argues that "no such naturalism can accommodate the sense of inescapable practical authority with which moral claims appear to be imbued".[66]

> What the moral naturalist evidently needs is a substantive and naturalizable account of "correct practical reasoning" [...] according to which any person, irrespective of her starting desires, would through such reasoning converge on certain practical conclusions that are broadly in line with what we would expect of moral requirements. [...] But no such adequate account exists.[67]

The issue of the inescapability of moral judgements will be taken up in more detail in a later section. At present it suffices to show that this claim, coupled with Joyce's epistemological argument calling the existence of moral facts into question, is what establishes Joyce's error theory. An evolutionary aetiology undermines any *prima facie* justification for our belief in moral facts, though it does not prove that they do *not* exist. When this agnosticism is coupled with the metaphysical oddness which such facts would seem to display, however, it can be seen that there is good reason to doubt their existence. The commitment of ordinary moral discourse to the existence of such facts is, however, non-negotiable. Ordinary moral discourse is therefore committed to an error.

It is worth noting at this point that Joyce is by no means the only philosopher to have advocated an evolutionary form of moral scepticism. Michael Ruse, for example, arrives at a conclusion very similar to Joyce's. He writes that:

> [s]ubstantive morality is a kind of illusion, put in place by our genes, in order to make us good social co-operators [...]. [T]he reason why the illusion is such a successful adaptation is that not only do we believe in substantive morality, but we also believe that substantive morality does have an objective foundation. [...] There are in fact no foundations.[68]

Sharon Street also argues that an evolutionary aetiology of morality should undermine our epistemic confidence in our moral judgements. Street claims that even if the truth of moral realism is granted, natural selection is not the kind of process likely to produce true beliefs about moral facts. That is, natural selection is not truth-tracking with regard to moral facts. It is unlikely to be the

---

[64] Joyce (2006) pp. 209 – 210.
[65] Joyce (2006) p. 209.
[66] Joyce (2006) p. 190.
[67] Joyce (2006) pp. 196 – 197.
[68] Ruse (2006) p. 21.

case, she argues, that a realm of independently existing moral facts would just *happen* to recommend those actions that make social life for creatures like us possible. Why, she asks, is a duty to care for one's children more likely to be a normative truth than "the judgment that infanticide is laudable [...], [or] the judgment that the fact that something is purple is a reason to scream at it"?[69] It is much more likely, she argues, that we have the moral beliefs we find ourselves with because it was evolutionarily adaptive for us to develop them.

> Of course it's possible that as a matter of sheer chance, some large portion of our evaluative judgments ended up true, due to a happy coincidence between the realist's independent evaluative truths and the evaluative direction in which natural selection tended to push us, but this would require a fluke of luck that's not only extremely unlikely, in view of the huge universe of logically possible evaluative judgments and truths, but also astoundingly convenient to the realist.[70]

A naturalist realist response to Joyce faces two challenges. The first of these will be to provide a reply to Joyce's claim that such a theory cannot successfully accommodate the categorical force of moral judgements. Even if such a response can be made, however, it will at best show that it is unreasonable to conclude that moral facts *cannot* exist. In light of Joyce's evolutionary aetiology, such a response will not provide sufficient reason to believe that moral facts *do actually* exist. The second challenge faced by the realist will therefore be to articulate a convincing account of the nature of those moral facts; furthermore, this must be an account whose explanatory force cannot be undermined by an evolutionary aetiology. This attempt will be the focus of subsequent chapters. Before discussing the categorical nature of moral judgements in more detail, however, I briefly consider three preliminary critical responses to Joyce. None of these, I argue, successfully undermines Joyce's argument for an error theory. Yet consideration of the third such response (in §6.iii) points to a promising way to accommodate Joyce's characterisation of the categorical force of moral judgements, which I then discuss in §7.

6. Preliminary Replies to Joyce

(i) Does Joyce's Argument Commit the Genetic Fallacy?

At first glance, Joyce's argument may seem to commit the genetic fallacy. This is the fallacy of claiming that the causal history of a belief shows that belief to be false. Note that such a move continues to be a commission of the genetic fallacy even in cases, like the evolutionary origins of morality, where the causal history of a belief is not a reliable indicator of its truth. Kahane provides an excellent example of an instance of the genetic fallacy; specifically:

> the suggestion that Marx's views on alienation have their source in the fact that he suffered from *hidradenitis suppurativa*, an agonizing skin disease said to cause self-

[69] Street (2006) p. 133.
[70] Street (2006) p. 122.

> loathing. [...] [Even if this were true, in order to] reject Marx's claims about alienation, we need to find flaws in the *reasons* he gave for them.[71]

In fact, however, Joyce does not commit the genetic fallacy. This is because he nowhere claims that his evolutionary aetiology proves moral judgements to be universally *false*. Rather, he claims that their status as seemingly objective *truths* is rendered epistemologically problematic. This more moderate claim allows for the possibility that our moral beliefs are in fact true. It simply states that, at present, we do not have legitimate grounds for believing this to be the case. Joyce's evolutionary aetiology should not therefore be read as an argument about the truth or falsity of our moral beliefs; rather, it should be read as an argument about the extent to which they are epistemologically justifiable. Understood as such, Joyce's position does not commit the genetic fallacy.

(ii) Objectivity Alternately Conceived

One strategy for devising a realist response to Joyce is to downplay the importance to moral discourse of the phenomenon about which we are supposedly in error. Such a strategy accepts a realist interpretation of moral discourse, but denies that moral objectivity need be as described by the error theorist. One such strategy is defended by David Phillips.

Phillips argues that Joyce's rejection of the categorical force of moral judgements is formally identical with Bernard William's rejection of external reasons for action. Phillips believes that he has an argument against Williams, and so it follows that if he is correct, Joyce's position will be similarly threatened. Before turning to Phillips' argument, however, a brief summary of Williams' scepticism about external reasons and its relation to Joyce's position is in order.

Williams held that an agent will be motivated to perform some action only if she believes that, by so doing, she will either satisfy one of her occurrent desires or achieve (or come closer to achieving) some pre-existing goal. Of course, she may be mistaken about the means of attaining a goal, or of satisfying a particular desire, and this may lead her to act in ways which thwart those aims. Nevertheless, her motivation will still be a result of her wanting to attain a goal or satisfy a desire. Thus, an agent will only consider herself to have a reason for action in the event that the action in question is instrumentally related to one of her subjectively held ends. Williams termed such reasons "internal reasons", contrasting them with "external reasons". The latter reasons also purport to give an agent a motivating reason to act, and to do so *irrespective* of any of her subjectively held ends. Yet Williams denied that external reasons were motivating reasons. That is, an agent will not take herself to have a reason for performing some action if the only reasons she has for so doing are external ones. Accordingly, only internal reasons are capable of explaining an agent's actions.

We may often believe that it is possible to provide an agent with a genuinely motivating external reason for action. Yet according to Williams this is a mistaken intuition. When we think that

---

[71] Kahane (2011) p. 105.

we have identified a motivating external reason, we have in fact covertly attributed a desire sensitive to that reason to the agent in question. We have, for example, assumed that an agent simply *must* desire to be healthy in later life, and that she will *therefore* be capable of being motivated to stop smoking, even if she does not now desire to do so. For Williams, any such attribution is illicit:

> external reason statements, when definitely isolated as such, are false, or incoherent, or really something else misleadingly expressed. [...] Those who use these words often seem, rather, to be entertaining an optimistic internal reason claim.[72]

Joyce does in fact characterise categorical reasons in terms of Williams' external reasons. Yet unlike Williams, Joyce does not deny the motivational potential of external reasons *tout court*. Rather, he claims that institutional norms such as the rules of etiquette constitute genuinely motivating external reasons for action. These, however, Joyce describes as "weak categorical imperatives".[73] These institutional norms are not such that they apply to an agent irrespective of her other, *internal* reasons, however.

> Etiquette may demand, categorically, that I do not speak with my mouth full, whether I wish to or not, but I only have a *reason* to refrain from speaking with my mouth full if I have some independent reason to be following the dictates of etiquette.[74]

Internal reasons therefore have the ability to "trump" external (i.e. institutional) reasons when certain conditions obtain. At such times, an agent's internal reasons take on a genuinely normative status; they represent what the agent ought to do all things considered.

Joyce goes on to explain that his objection is to "external reasons claims that do not know their place – that overstep themselves by claiming to transcend all institutions. Such, I have argued, are moral reasons".[75] That is to say, moral reasons are in fact institution-dependent external reasons. As such, they may, according to Joyce, legitimately be trumped by an agent's internal reasons, should these conflict with those moral reasons. In such cases, it is therefore incorrect to describe immoral actions as practically irrational. Thus, Joyce claims that:

> there would be no problem with morality and its reasons if only it presented itself as an institution. But it doesn't. It presents itself as something with ubiquitous and inescapable authority.[76]

Phillips rejects this position, denying that internal reasons can constitute genuinely normative reasons. He argues that a genuine reason for action cannot be derived from an internal reason, as internal reasons are contingent upon what Williams terms an agent's "subjective motivational set".[77] Thus Phillips argues that according to Joyce and Williams, "genuine reasons are grounded in agent's desires. The desires are the source, the only source, of genuine practical

---

[72] Williams (1981) p. 111.
[73] Joyce (2001) p. 36.
[74] Joyce (2001) p. 37. Original emphasis.
[75] Joyce (2001) p. 133.
[76] Joyce (2001) p. 104.
[77] Williams (1981) p. 102. Emphasis removed.

reasons".[78] This is a problem, however, as the contents of an agent's subjective motivational set are, "to a significant extent, [...] arbitrary, contingent fact[s] about her. Why are her desires therefore a source, or the only source, of authority?"[79] According to Phillips, that is, an agent's desires are just not the sort of things which are capable of generating genuinely normative reasons for action. This would make normativity much too "arbitrary and contingent".[80]

Furthermore, Phillips claims, the Williams-Joyce position:

> conflicts with firm intuitions about the reasons we have. [...] [S]uppose that Alice wants to smoke cigarettes. She does not enjoy smoking cigarettes. She knows about the deleterious consequences for her health. But her desire to smoke cigarettes would, for all that, survive the kind of procedurally rational reflection which William's view allows. He is then compelled to say that Alice has a reason to smoke cigarettes. But surely the mere fact that Alice has this desire [...] does not show that she has a good reason to smoke cigarettes.[81]

Phillips' claim that the Williams-Joyce position conflicts with our intuitions is true. Still, this does not by itself show that position to be false. Our intuitions may well deceive us. Furthermore, Phillips' scenario is not entirely convincing. The fact that Alice continues to smoke shows that she is for *some* reason motivated to do so. Perhaps she thinks that doing so makes her look cool. Of course, Phillips can claim that this does not constitute a *normative* reason. Alice has more reason to be concerned about her health than to look cool. Certainly this thought has some intuitive pull. After all, doesn't *everyone* care more about their health than about looking cool? But what if Alice does not? What if she really would prefer to look cool than to be healthy? In that event, we have fallen into the psychological trap which Williams cautioned against: that of illicitly attributing internal reasons to an agent. If Alice's reason for continuing to smoke is more important to her than any other reasons which count in favour of quitting, then it is by no means clear how she can be said to be acting in a practically irrational way. To insist that this simply must be the case is begging the question, at least in the absence of an alternative account of normative reasons. Yet Phillips believes that he can provide just such an account.

Phillips argues that the putatively normative internal reasons endorsed by Williams and Joyce are hypothetical imperatives. This is because "[s]uch imperatives bind only on a certain hypothesis, the hypothesis involved in having a desire, that a certain end is good."[82] Phillips goes on to argue that these hypothetical imperatives need not be derived solely from the subjective motivational sets of individual agents. Rather, "[s]uch hypotheses may derive from presupposed frameworks, institutions, or contexts, which are not the products of the desires of the agent".[83] Phillips suggests that these imperatives are preferable candidates for the role of normative reasons, as they do not rely on "an arbitrary privileging of the agent's own current desires".[84]

---

[78] Phillips (2010) p. 97.
[79] Phillips (2010) p. 91.
[80] Phillips (2010) p. 98.
[81] Phillips (2010) p. 91.
[82] Phillips (2010) p. 97.
[83] Phillips (2010) p. 97.
[84] Phillips (2010) p. 97.

If successful, Phillips' account provides an alternative to Joyce's conception of normativity. As a result, Joyce's argument for an error theory would be blocked, as genuine normative reasons for action would exist in the form of institution-dependent external reasons. These, it will be recalled, were found to be unproblematic by Joyce. Thus Phillips will have shown that moral objectivity is different from Joyce's conception of it, and that, properly conceived, it is unproblematic even by Joyce's own lights.

In fact, however, Phillips' argument does not succeed. This is because it remains deeply unclear how a presupposed framework can, by its mere existence, provide an agent with a normative reason for action. The frameworks which Phillips has in mind, it is here assumed, are socially constructed. As such, they can be construed as goals or desires which are shared by a large percentage of a society. But why should the fact that an end is shared by a group of individuals provide an agent with a normative reason to adopt that end herself? Surely it cannot. Suppose some framework provides a putatively normative, hypothetical-external reason to for an agent to ϕ. Further suppose that an agent finds herself in a situation in which to ϕ would massively disadvantage her. Why, then, should she ϕ? It is not enough simply to reiterate that the framework demands that she do so: she already knows that *according to that framework* she ought to ϕ. The question is why she should continue to deliberate about ϕ-ing from within that framework. Phillips has no response to this question. Indeed, he admits as much when he notes that such frameworks are not "straightforwardly or presupposition-independently correct. In that sense their authority is incomplete."[85]

Furthermore, Phillips has not shown that his account of normative reasons makes them significantly less arbitrary than does the Williams-Joyce account. If the frameworks of which Phillips writes are indeed the product of communally shared goals or desires, but the desires of the agents comprising that community are problematically arbitrary (too arbitrary, that is, for them to generate genuinely normative reasons), then why should the framework generated by the agglomeration of those goals and desires not *itself* be problematically arbitrary?

Phillips' account of normative reasons, then, is not a success. It makes them no more rationally binding, and no less problematically arbitrary, than Phillips himself takes them to be on the Joyce-Williams account.

(iii) Should Metaphysics Make a Difference to Normative Judgement?

Is it possible to accept everything which Joyce says about the nature of moral discourse, the evolutionary aetiology of our moral judgements, and the difficulty of giving a satisfactory naturalistic account of moral properties, and yet to reject the claim that these considerations make our pre-theoretical moral judgements problematic? This may seem like a tall order, but such a position has in fact been argued for by Jamie Dreier. Dreier argues against error theory by developing an analogy between the relation of metaethics to ordinary moral discourse, and the relation of physics to pre-theoretical, or "folk", discourse about physics. Dreier claims that:

---

[85] Phillips (2010) p. 98.

> [a] sophisticated understanding of the physics which underlies these features of the world [i.e. solidity and simultaneity] shows that they are not at all what we think they are. Solid objects do not uniformly fill the space they occupy with undifferentiated matter; rather, they are mostly space, they are (constituted by) crystalline lattices of atoms held together by electromagnetic forces. [...] But sophisticated physics does not tell us that there are no solid objects [...]. The theoretic gloss on our ordinary judgments [about physics] is not properly understood as infecting every first-order judgment [about physics] with false presupposition. Why, then, should false metaphysics infect our moralizing?[86]

Dreier's argument in this passage is directed at Mackie's argument for an error theory. Given the similarities between Joyce's position and Mackie's, however, it can be taken to apply to Joyce's version of error theory in equal measure.

The argument can be summarised as follows. Just as our experience of tables and chairs guarantees the factuality of their solidity (indeed, *makes it the case* that they are solid objects), so too, Dreier suggests, does our experience of moral judgements as objectively true or false guarantee that they are such. It is simply that what *makes* them true or false is something other than previously thought. Dreier doubts Mackie's claim that his argument shows there to be nothing which is genuinely morally right or wrong. Rather, Dreier suggests, what Mackie has actually shown is that:

> plenty of things (those things we ordinarily think are morally right and wrong) really are, only the fact that they are is partly constituted by our thinking so, or by our sentiments [...].[87]

In order to assess the merit of Dreier's argument, it is instructive to begin by considering one of the ways in which error in folk-physics seems different to error in ordinary moral discourse. This is that in the case of folk-physics, we do not tend to react to the discovery of such error in the same way. Thus, upon hearing that solid objects are in fact mostly empty space, one may well act with surprise, and perhaps disbelief; one would typically not, however, be scandalised by such a claim.[88] This is not the case when dealing with ordinary moral discourse. The claim that such discourse is systematically in error typically meets with much more outright hostility than do comparable claims about folk-physics. If such error has no bearing on the truth of our first-order moral judgements, however, why should this be so? Are we simply mistaken to react differently to error in ordinary moral discourse than we do to error in folk-physics?

It is possible to develop an answer to this question by drawing a parallel with debates over aesthetic value. Suppose Anna, an art lover, buys a Cezanne. She admires it daily, and derives much pleasure from doing so. One day, Anna's friend Beth, who happens to be an art historian, pays a visit in order to admire Anna's purchase. After a careful inspection Beth regretfully informs Anna that the

---

[86] Dreier (2010) p. 79.

[87] Dreier (2010) pp. 78 – 79.

[88] It might be thought that there are historical counterexamples which undermine this claim, one such being resistance to the heliocentric model of the solar system. However, as I interpret it, the outrage which met the heliocentric model issued from a perceived slight to religious, and thus by extension moral, beliefs associated with mankind's place in the universe. The source of the outrage was not the heliocentric theory *per se*, but its apparent threat to a particular religious paradigm.

Cezanne is a forgery. There are minor details, imperceptible to Anna's untrained eye, which allow Beth to identify it as such. Ever after, Anna is unable to derive quite the same amount of pleasure from the painting. Despite the fact that it looks the same to her as it did before, and the fact that she is unable to perceive the forgery-making properties which Beth assures her to be there, Anna feels that the painting has lost some of its aesthetic value.

This scenario is intended to parallel the devaluation of our first-order moral judgements, thought to follow from an increased understanding of the nature of such judgements. Anna's reaction here seems natural, yet it is also hard to explain: the painting is physically the same as it ever was, and Anna cannot perceive the properties which identify it as a forgery. Why, then, should Anna's enjoyment of her painting be adversely affected? Similarly, the phenomenal experience of moral judgements as objectively true or false is not altered by error theoretic arguments to the contrary. Why, then, should such conclusions be thought to undermine that objectivity? If it is possible to answer this question in the case of aesthetic value, perhaps that answer will generalise to cover moral value as well.

One possible answer to the aesthetic version of the question is to claim that aesthetic value is not solely determined by an object's physical composition. Such value is also generated by its status as a historical artefact. Thus Luise Morton and Thomas Foster argue that:

> knowledge of differences in the origins of any two indiscernible material objects will enable us to discriminate aesthetic differences between them when we perceive them *as works of art*.[89]

Analogously then, it might be argued that part of the value which we attach to our moral beliefs stems from our conception of their origins. Perhaps, for example, we take those beliefs to be the product of Divine law. To challenge that conception is to therefore challenge the beliefs themselves.

This move is problematic, however. Firstly, note that it is very close to being a straightforward commission of the genetic fallacy (see§6.i). Even if we *do* have a psychological tendency to react in this way, doing so is nevertheless irrational. If we were thinking clearly, then we would not feel morality to be so threatened. Of course, it may be the case that we are simply *unable* to think clearly about this issue; unable to prevent ourselves from committing some form of genetic fallacy. Even if this were true, however, it would not show that such devaluation could not be reversed. Why should an evolutionary aetiology such as Joyce's be any less capable of conferring value upon morality than the claim that it is the product of Divine law? When we consider the millions of years during which the human mind has been shaped by natural selection, and all the contingent events which shaped that process, the results can seem truly precious. Why should this not be enough to confer as much value upon morality as any other genealogical narrative? Surely, it should be. But if this is true, then the reply to Dreier suggested by Morton and Foster does not work.

There is a second line of response to Dreier yet to consider, however. When a forged piece of art loses some of its aesthetic value, perhaps it does so because it has been shown to be non-veridical. Forgeries purport to be other than what they are: this is what distinguishes them from copies, imitations, or works "in the style of". Bernard Williams has argued that truth possesses a

[89] Morton and Foster (1991) p. 156.

basic, unconditional value for humans.[90] Accordingly, the value which we attach to truth is what drives us to respond less positively to a forgery, no matter how perfect a copy, than we do to the original. In the same way, it is the value which we attach to truth which undermines the objective truth and falsity of moral judgements in the light of Joyce's genealogy.

According to Joyce, moral judgements are categorical. That is, we take them to possess non-instrumental validity. When we value the truth of those judgements in a naïve frame of mind, therefore, we value it *qua* non-instrumental truth. Yet what Joyce's account claims to show is that moral judgements are only true instrumentally. To the extent that we accept Joyce' arguments we value, *qua* putative truth, our newfound knowledge that moral judgements are only instrumentally valid. This sets up a psychological conflict between the value that we ascribe to putatively non-instrumental moral truths, and the value that we ascribe to metaethical truths about the instrumental nature of moral discourse. Williams expresses the same thought when he writes that:

> a truthful historical account is likely to reveal a radical contingency in our current ethical conceptions. [...] This sense of contingency is likely to be in tension with something that our ethical ideas themselves demand, a recognition of their authority. The tension here is made worse by a feature of modern ethical systems, that they try to combine authority with transparency, and in aiming to be transparent – an aim that is part of their special concern with truthfulness – they encourage reflection on themselves in a style that reveals their contingency.[91]

This argument shows that Dreier is mistaken in claiming that an error theory of moral discourse need have no repercussions for the way in which we evaluate the truth of first-order moral judgements. Yet Dreier's position can be adapted to provide the basis of a satisfying, naturalistic account of the categorical force of moral judgements. It is to this account that I now turn.

7. The Nature of Categorical Judgements

In this section I argue not only that the concept of categorical moral reasons can be cashed out naturalistically, but also that this can be achieved from within the Williams-Joyce perspective of reasons internalism. Doing so shows that moral facts need not be as metaphysically queer as Joyce and Mackie suppose, and so significantly reduces the force of their sceptical challenge.

(i) Categorical Moral Reasons and Motivation

To begin with, it is important to emphasise what an account of categorical moral reasons need *not* do. Firstly, and quite un-controversially, moral reasons need not exert any motivational

[90] In Williams (2002).
[91] Williams (2002) pp. 20 – 21.

influence whatsoever on an agent *who does not accept them as such*. Jack may tell Jill, quite in earnest, that it is morally obligatory to torture innocent children. But of course his doing so need not and indeed *ought* not to motivate Jill to abduct and torture a child. More likely it will motivate Jill to avoid associating with Jack, and perhaps also to suggest that he see a psychiatrist. Jack's moral compass, she might well assume, is badly misaligned.

Secondly, and somewhat more controversially, moral reasons need not motivate an agent *even when that agent recognises them as such*. This point can be argued for most clearly by considering the perennial philosophical problem of the immoralist. For Joyce, the immoralist is personified in the Platonic representation of Gyges, "a Lydian shepherd [...] who comes across a ring of invisibility, which he uses in morally objectionable ways to gain whatever he wishes".[92] For the purposes of the present discussion, however, I will refer to the immoralist *par excellence*: John Milton's Satan.

For Satan, it is "Better to reign in Hell, than serve in Heav'n".[93] Later, as a result of his exile from Heaven, Satan famously proclaims "Evil be thou my good; by thee at least/Divided empire with Heav'n's King I hold".[94] Now, it is safe to say that Satan knows which actions are morally good and which ones are not. Furthermore, he knows *why* they are morally good and bad (i.e. by God's decree), and that they *truly are* morally good and bad (as Satan is personally acquainted with their Author, he can have no doubts about God's existence). Satan, then, has perfect moral knowledge. He also knows, we may safely assume, that morality is categorically binding: presumably God did not decree that one ought never to ϕ, unless it happens to be the case that one is engaged in a rebellious campaign against Him, in which case it is morally permissible to ϕ. Thus, given his subjective motivational set, Satan's justified true belief that to ϕ in circumstances C is morally forbidden *gives Satan a reason always to ϕ in circumstances C*. Thus, even in the event that there are categorically normative facts, and that an agent knows what these facts are, that agent need not necessarily have a reason to act in accordance with them.

Under what conditions, then, *does* an agent have a motivating reason to act in accordance with a normative fact? And how can such reasons be said to be categorically binding if they are subject to that agent's having a particular subjective motivational set?

The answer to the first of these questions is perfectly straightforward. An agent has a reason to act in accordance with a moral imperative if doing so satisfies one of the desires, or promotes one of the ends, in her subjective motivational set. Note, however, that this does *not* mean that a motivation to act in accordance with a moral reason will only come about when acting morally is instrumental to attaining some other, non-moral, end. Williams cautions his readers that, although he typically characterises the notion of a subjective motivational set

> in terms of desires [...] [it] can contain such things as dispositions of evaluation, patterns of emotional reaction, personal loyalties, and various projects [...] embodying commitments of the agent.[95]

---

[92] Joyce (2001) p. 32.
[93] Milton (2000) i: 263.
[94] Milton (2000) iv: 110 – 111.
[95] Williams (1981) p. 105.

That is, an agent may be motivated to act morally simply because she is committed to doing so.

What, then, of the categorical force of moral judgements? As will be recalled from the previous section, Dreier suggests that:

> plenty of things (those things we ordinarily think are morally right and wrong) really are, only the fact that they are is partly constituted by our thinking so, or by our sentiments [...].[96]

Though this line of thought was unsuccessful as an argument for the objective truth or falsity of moral judgement, it fares much better when used to naturalise the categorical *force* of morality. Thus, in brief, a moral judgement is categorically binding on an agent whenever she *experiences* that judgement as categorically binding.

At first glance, this may not seem to vouchsafe anything like the sort of categoricity that Joyce has in mind. For example, he would no doubt wish to re-emphasise his claim that ordinary moral discourse purports to give an agent categorical reasons for action *even if that agent is unmoved by them*. It simply will not do to say that an agent need not have refrained from committing murder simply because she did not judge that, all things considered, she *ought* so to refrain.

In addition, Joyce would resist any attempt to characterise the subjective motivational set of every agent as containing a desire to act morally:

> To present the-good-of-fellows as the object of a *desire* which all humans have (reliably, on all occasions), a desire from which no human could escape, is to divest the [natural sympathy] thesis of credibility. We're all too familiar with counter-examples.[97]

Yet when the details are filled in, the account which I propose can be seen to meet these challenges.

(ii) Moral Beliefs and Categorical Moral Judgements

It will be recalled that, according to Joyce's account, "moral judgments (as speech acts) express both beliefs and conative non-belief states".[98] In what follows, I will refer to Joyce's account of moral judgements as a "hybrid" account. Joyce denies that his view commits him to the existence of "besires": single mental states which have some of the features of desire, and some of the features of belief. Thus he claims that:

> [t]he fact that a kind of utterance may express belief *and* attitude implies nothing about the modal relations holding between the speaker's belief states and attitude states. This worry deflected, there seems nothing philosophically troubling in the idea of a linguistic

[96] Dreier (2010) pp. 78 – 79.
[97] Joyce (2001) p. 61.
[98] Joyce (2006) p. 56. For a related view see Copp (2001; 2009).

convention that decrees that when S is uttered in C the speaker thereby expresses *two* mental states.[99]

Joyce's arguments in support of his hybrid account of moral judgement (for which, see §4.i) are persuasive, and I will not dispute them. In fact, there is an additional reason to prefer Joyce's account of moral judgement to those accounts which see moral judgements as *either* cognitive *or* conative: Joyce's position actually makes it easier to meet the objections which he makes against attempts to naturalise moral properties. This is because a hybrid account of moral judgement is able to accommodate intuitions defended by internalists about moral motivation, as well as conflicting intuitions defended by externalists about moral motivation. Specifically, it makes it possible to explain how moral judgements can be described as intrinsically motivational (satisfying Joyce's claim that they have an essentially practical, action-guiding aspect), whilst *denying* that moral *facts* must themselves be motivational (as this would make them metaphysically queer). In the rest of this subsection, I show how this can be done. I then show how such an account makes it possible to cash out the categorical force of moral judgement in a strictly naturalistic way.

To begin with, it should be noted that it follows from Joyce's account that it is logically possible for an agent to have a moral belief, yet not to possess a particular conative *attitude* towards that belief. This follows from the distinct psychological existences of the cognitive and conative aspects of moral judgement. But if making a moral judgement expresses a belief state *and* a conative non-belief state, then an agent who finds herself without the requisite conative attitude *is by definition not making a moral judgement*.

What, then, is such an agent doing? She is engaged in what I will refer to as "moralising". Moralising, as I intend the term to be understood, is a ubiquitous phenomenon in moral philosophy. Put simply, it is emotionally disengaged moral thought. Because it is emotionally disengaged, however, the moral beliefs arrived at by engaging in moralising are motivationally inert. A die-hard utilitarian moralises, for example, when she says: "torturing an innocent child may be the morally right thing to do in some extreme circumstances; but I could never bring myself to torture anyone. I just haven't got it in me." Here, the utilitarian has arrived at a moralistic conclusion, and has clearly formed a moral *belief*, but she has *not* made a genuine moral *judgement*. If that were the case, she would find herself motivated to torture an innocent child, at least in the event that morality should demand it of her. Another famous example of moralising is provided by Michael Smith:

> Suppose we are sitting together one Sunday afternoon. World Vision is out collecting money for famine relief, so we are waiting to hear a knock on the door. [...] We debate the pros and cons of contributing and, let's suppose, after some discussion, you convince me that I should contribute.[100]

According to Smith's scenario, he nevertheless fails to be motivated to give money to charity when World Vision arrives at the door. This lack of motivation is what identifies the discussion in which Smith was previously engaged as an instance of moralising. That discussion resulted in Smith acquiring a true (let us suppose) moral belief, but it did not cause him to make a corresponding

[99] Joyce (2006) p. 57.
[100] Smith (1994) p. 6.

moral *judgement*. Making such judgments is not simply a matter of *believing* that one ought to give to charity; it is also a question of having an appropriate attitude *towards* that belief.

Drawing a distinction between moral belief and moral judgement in this way makes it possible to accommodate elements of both internalist and externalist views on moral motivation. This is because such a distinction is often overlooked, with the terms "belief" and "judgement" being used interchangeably. Thus Adina Roskies writes that:

> Internalist theses are sometimes couched in terms of belief, and other times in terms of judgment. Since judgment plausibly entails belief, I take it that if I provide arguments sufficient to refute belief-internalism, these also, *a fortiori* refute judgment-internalism.[101]

The claim which externalists about motivation wish to make secure is that there is only a contingent connection between sincere moral belief and moral motivation. They therefore reject the internalist's view that, when one sincerely believes there to be a moral reason to ϕ, one is *ceteris paribus* motivated to ϕ. Roskies uses the example of acquired sociopathy to this end. She shows that damage to specific areas of the brain results in a loss of moral motivation: individuals who suffer from such brain damage "all appear to have particular difficulty in acting in accordance with social mores".[102] Yet these individuals do not simply suffer from a loss of moral knowledge (or a change in their moral beliefs).

> Before their brain damage, [...] subjects clearly had mastery of moral terms, for they were just like you or me. [...] According to all tests, their language and declarative knowledge structures [remain] intact. [...] [T]hus, if they initially knew the meaning of moral terms, they still do.[103]

Rather, this form of brain damage affects "the neural systems for arousal and emotion",[104] and sufferers "display and report attenuated or absence of affect when faced with situations that reliably elicit emotions".[105] Roskies argues that this shows there to be no necessary connection between sincere moral belief and moral motivation.

A theory according to which moral judgements express both beliefs and conative attitudes can accommodate Roskies' argument, and accept that there is no necessary connection between moral belief and moral motivation. This is precisely because, as Roskies argues, in cases like that of the acquired sociopath, the conative aspect of moral judgement is absent. Yet, by insisting on drawing a firm distinction between moral judgement and moral *belief*, such a theory can also accommodate the internalist's claim that moral judgements are motivational. This is because such judgements are, constitutively, partly conative. Thus it is the conative aspect of moral judgement from which motivation derives, and not the cognitive aspect.

---

[101] Roskies (2003) p. 53.
[102] Roskies (2003) p. 56.
[103] Roskies (2003) p. 60.
[104] Roskies (2003) p. 55.
[105] Roskies (2003) p. 57.

The foregoing analysis helps to ground the observation, made with reference to Satan in §7.i, that even when true, moral beliefs are not in themselves motivational. However, true moral beliefs are simply true beliefs about moral *properties.* It therefore follows that perceiving or recognising such properties is not sufficient to motivate moral behaviour. Moral facts, like non-moral facts, are motivationally inert. Joyce's complaint that naturalism is unable to provide a satisfactory account of how moral properties could be intrinsically motivational therefore turns out not to be a problem for naturalism: no such account need actually be given. It should be noted that this perspective sits comfortably with Joyce's commitment to reasons internalism:[106] the motivating, conative aspect of moral judgement can simply be said to be part of an agent's subjective motivational set. Additionally, and importantly, the conceptual shift away from motivating-moral-facts/properties and towards motivating-moral-*judgements* takes a lot of the queerness out of moral facts. It will be recalled that Mackie's objection to the queerness of moral properties was based on their intrinsically motivational status. This made them uniquely unlike any other natural properties. As this discussion has shown, Mackie was mistaken to claim that this was a necessary feature of moral properties. They are therefore much less metaphysically queer than he supposed.

Now, a plausible way of construing the putatively categorical force of moral judgements is to say that they exert a stronger motivational influence on our actions than non-moral judgements. Given this, it is reasonable to construe categoricity as a further, related feature of that agent's motivational set. It will be recalled that Joyce's scepticism about the possibility of naturalising categoricity was directed at categorical moral *facts*. Yet intrinsically motivational facts are no longer in need of a naturalistic explanation. Rather, what is required is a naturalistic account of a particular feature of an agent's subjective motivational set. Clearly, this is much less of a burden for the naturalist than it would be to provide an account of intrinsically categorically motivating moral properties. I will now sketch what a naturalistic account of a categorically demanding feature of an agent's subjective motivational set might look like.

According to my argument, adapted from Dreier's position, when moral judgements are phenomenally experienced as categorically binding, this experience is enough to guarantee that they are legitimately describable as such. Morality's categorical force is a phenomenal quality of moral judgement, not an independent property in need of further explanation. When moral judgments are felt to be categorically demanding, this is simply in virtue of their constitutive conative element engaging our affects more forcefully than any of the other contents of our subjective motivational set. When deliberating about how to act, moral judgements simply matter more to us than our other, non-moral goals. Joyce claims that:

> [t]he term "categorical imperative" [...] captures a quite simple mental [...] phenomenon. If a person thinks, concerning, say, the help of her child, "I just have to" [...] then she is wielding a categorical imperative.

He is right about this. I would simply add that she is wielding that imperative correctly.

Note that it does not follow from this account that we will, necessarily, always do what we judge we morally ought to do. This would be a highly implausible conclusion, and would deny the existence, and possibility, of akratic agency. All that I am claiming here is that, when we *do* act

---

[106] See §6.ii.

against our sincere moral judgements, then by our own lights we judge that we are not doing as we *ought* to do. Thus we typically find instances of akrasia to be a cause for regret or remorse. Indeed, so forcefully do we judge in favour of acting morally, that we typically regret failing to do so even when this is a result of holding false moral beliefs. That is, we regret failing to act morally even when we were trying to do so, but were misguided in the attempt.

My interpretation of morality's categorical force can be supported by appealing to its evolutionary development. The problem of social cooperation which morality evolved to solve was deeply pressing, presenting our forebears with a great many opportunities to selfishly exploit their conspecifics. It was not in their genetic interests to do so, however; it was far better (i.e. genetically fitter) to become more able to resist the temptation to betray those partners with whom they were cooperating. Accordingly, there was selection pressure for the psychological cost of such betrayal to become increasingly greater, the better to act as a disincentive against cheating. This process gradually occurred, finally bringing it about that cheating came to be experienced as *simply* forbidden, and harmonious cooperation as *simply* required.

This is not to suggest, however, that natural selection has instilled in us a "desire for the good of others",[107] for the sake of whose satisfaction all of our moral actions are performed. As noted above, Joyce finds such a notion deeply implausible, and rightly so. Rather, those actions which we take to be moral are valued non-instrumentally, as a result of our "dispositions of evaluation [and] patterns of emotional reaction".[108] Moral emotions such as guilt and shame are experienced as intrinsically unpleasant, and so we avoid acting in ways which arouse these feelings in us. Conversely, those emotions produced by the moral approbation of others are intrinsically pleasant, and encourage us to act in ways which encourage such approbation.

Joyce describes categorical judgements as judgements which an agent "cannot legitimately ignore";[109] they have "authority over people *irrespective of their interests*".[110] Both of these statements are true of moral judgements considered as phenomenally categorical. An agent who acts in opposition to her moral judgement is, by her own lights, acting in a practically irrational way: she is ignoring the feature of her subjective motivational set about which she herself cares the most. The interpretation I propose is therefore in accordance with ordinary moral discourse.

But what about the psychology of the evil agent? What about Satan? The account which I have just sketched still leaves us unable to provide Satan with a reason to change his ways. Does it, for that reason, fail as an account of categorical moral demands? Unsurprisingly, I do not believe that it does.

Satan can be seen as an infallible moraliser. He possesses nothing but true moral beliefs. Yet he lacks a conative disposition towards acting in accordance with those beliefs. This is not to say that he lacks any conative disposition whatsoever, of course. Indeed, his conative disposition strongly motivates him to act in ways which he knows to be *im*moral. Satan *does* make moral judgements, then. It is just that those moral judgements are *evil*.

---

[107] Joyce (2001) p. 60.
[108] Williams (1981) p. 105.
[109] Joyce (2001) p. 41.
[110] Joyce (2006) p. 196.

Of course, we typically do not encounter agents as exotic as Milton's Satan in our daily lives. Nevertheless, Joyce argues that moral discourse is committed to the existence of reasons that could motivate Satan to act in morally good ways, and that the example of such an evil agent shows there to be no such reasons. There are two points to be made in relation to this claim. Firstly, why should we say that *reasons* to act differently are what Satan is in need of? It is not that Satan believes murder, torture and other misdeeds to be morally good: were that the case, he could be talked out of pursuing them by a sufficiently persuasive philosophical argument defending some other conception of "morally good". In the current scenario, however, the object of our condemnation is not Satan's *rationality*; it is his lust for evil. Satan's perverse psychology need not put him outside the scope of our moral condemnation, however. It is perfectly permissible for us to say that he is acting badly, as the truth of one's moral beliefs is independent of the extent to which one is motivated by them. Indeed, Satan would wholeheartedly agree with the judgement that he acts badly. The problem (for us) is that he has no desire to act *well*. Unless there is something in Satan's subjective motivational set which will allow us to show him how acting morally will realise something which he values, there is nothing we can say that will give him a reason to change his evil ways. Thus, as Nick Zangwill explains:

> the requirements of morality bear on us regardless of our desires. But we should be wary of putting this in terms of "reasons for action". [...] The amoralist [...] would be instrumentally irrational to heed moral requirements. In *that* sense, he has *no* reason to be moral. But the requirements of morality still apply to him. [...] [T]he amoralist is subject to a moral requirement but not a rational requirement to do what morality demands.[111]

The typical moral agent, however, subscribes to what James Doyle terms "moral commitment". Moral commitment is, for Doyle, "an essential part of the moral point of view".[112] Put simply, to be morally committed is to refuse to consider moral constraints upon one's actions to be practically defeasible. That is, the morally committed agent is one who "[refuses] to contemplate with any seriousness the supposed possibility that these [moral] constraints might not really be binding".[113] Amoral agents like Satan cannot rationally be persuaded to mend their wicked ways, simply because:

> the acquisition of a new belief, such as might occur as the result of being persuaded by an argument, is incapable of bringing about the change in the fundamental orientation of one's life which moral commitment, as an essential feature of the moral point of view, involves.[114]

This brings us to the second point relevant to discussions of evil agents. Quite simply, it is that most people are not like Satan. The vast majority of people *do* have a desire to act morally; they are highly sensitive to the criticism that their actions are immoral, and will be moved to justify them against such a charge. That is, they will agree that some things are morally forbidden, or required.

---

[111] Zangwill (2003) p. 152.
[112] Doyle (2000) p. 11.
[113] Doyle (2000) p. 10.
[114] Doyle (2000) pp. 21 – 22.

For such agents, the contentious issue is not *whether or not* there are categorical moral demands; it is one of identifying *which actions* are subject to them.

Appealing to the a-typicality of Satan's psychology may seem like a flagrant attempt at evading Joyce's challenge. This is not what I mean to do, however. The a-typicality of evil agency in fact threatens the coherence of Joyce's conception of the function of categorical reasons in ordinary moral discourse. To see this, we must ask: "why should ordinary moral discourse be such that it refers to all *logically possible* agents, as opposed to all (or the vast majority of) *actual* agents?" It is surely less plausible for the former to be the case, if for no other reason than that ordinary moral discourse hardly ever has to *address* such agents. Why should we accept that, whenever they make moral judgements, ordinary moral practitioners have in mind reasons which necessarily apply to every logically possible agent, irrespective of how their psychologies may differ from ours? Why should we even accept that ordinary moral practitioners can coherently *conceive* of the psychology of every logically possible agent?[115] If the chances of meeting a genuinely evil agent, as opposed to someone who is simply morally misguided, are sufficiently remote, why would moral discourse historically develop so as to *refer* to such evil agents? The most plausible answer is that it would not.

Joyce's conception of categorical moral reasons is therefore deeply implausible as an analysis of ordinary moral discourse. The alternative account which I have provided is able to accommodate the way in which ordinary moral practitioners use moral language, and does so without drawing on philosophically dubious claims about what such language is tacitly committed to.

In summary, then, categorical moral reasons cannot provide evil agents like Satan with reasons to be moral. Yet this is not a problem for an analysis of the categorical force of ordinary moral judgements. It is unreasonable to expect *ordinary* moral discourse to purport to provide reasons for all logically possible agents, as opposed to those who are *ordinarily* encountered by moral practitioners. The account which I propose is therefore a defensible, naturalistic account of the categorical nature of moral judgement. The fact that it does not do everything which Joyce demands of a categorical reason is no objection to it: Joyce's demands are implausibly stringent.

Finally, it should be noted that one of the felicitous consequences of taking up the perspective on categorical moral reasons which I endorse is that it defuses Phillips' problem of contingency. If categorically valuing moral reasons is something which has been instilled in us by natural selection, then the worry that being moved by such reasons is dependent upon contingent features of an agent's subjective motivational set is significantly diminished. Of course, such contingency still obtains, but it is an evolutionary contingency, and not a will-she-won't-she, toss-of-the-coin contingency. That is, having the appropriate subjective motivational set is contingent in the way that having two eyes, or only one head, is contingent. These facts about us are contingent upon our being the sort of creatures which we are. Clearly, however, this sort of contingency is much less troubling for morality than are the contingent psychological (or societal) preferences and goals about which Phillips is concerned.

---

[115] This is not intended as a slight upon the faculties of the ordinary moral practitioner: the same question may legitimately be asked even of philosophers.

8. Conclusion

This chapter introduced Joyce's evolutionarily motivated argument for a moral error theory. The evolutionary narrative provided by Joyce is intended to overcome realist intuitions motivated by considerations of epistemological conservatism (§2). It does this by providing an aetiology of our moral beliefs which does not contain an appeal to the existence of normative facts (§3). Following the exposition of this aetiology, §4 gave a short summary of Joyce's philosophical account of ordinary moral discourse. In §5 it was shown how this account, when conjoined with Joyce's evolutionary aetiology, is able to support an argument for moral error theory. The argument's structure is essentially to claim, firstly, that evolutionary data casts serious doubt on the existence of moral facts; secondly, that a naturalistic approach to ethics is the only approach not delegitimised by an evolutionary perspective on ethics; and thirdly, that naturalism is unable to make sense of the philosophical commitments of ordinary moral discourse. §6 considered some preliminary replies to Joyce's argument, and argued that none of those replies was ultimately successful. However, the argument made by Dreier in §6.iii suggested a possible line of response to Joyce's scepticism regarding attempts "to naturalize moral clout".[116] This response was developed more fully in §7, where it was argued that the categorical force of moral judgements can be understood in terms of an overriding conative commitment to morally right action. Categoricity should be construed not as a fact about moral properties, but as feature of the psychology of moral agents.

This chapter has argued for the availability of a naturalistic account of categorical moral reasons. Yet it will be recalled that Joyce's evolutionary aetiology was not aimed solely at establishing this conclusion. His denial that categoricity could be made sense of was rather the icing on a sceptical epistemological cake. To revert to the subject of Joyce's belief pill analogy, his argument was that not only do we lack reliable evidence either for or against the occurrence of the battle of Waterloo, but that the occurrence of the battle of Waterloo is metaphysically problematic. We therefore have good reason, he argues, to doubt that it took place. My conceptual defence of categoricity has hopefully removed this metaphysical worry. In the terms of Joyce's analogy, the battle of Waterloo is the sort of thing that could have taken place in a naturalistic universe. Of course, to show that Waterloo *could* have been fought is not thereby to show that it *was* fought. Similarly, showing that categorical moral properties are metaphysically unproblematic is not thereby to show that they *exist*. Joyce's evolutionary aetiology is still able to explain the content of our moral judgements without making reference to the existence of moral facts. A complete realist response to Joyce will therefore show that moral facts merit inclusion in some such aetiology. That is, it will not only show that moral facts *could* exist; it will also show that we have reason to believe that they *do* exist. This will be the task of the following chapters. The first argument to be considered as a means to this end will be that made by Philippa Foot.

[116] Joyce (2006) p. 196.

Chapter Four

The Evolution of Virtue

1. Introduction

In this chapter I discuss Philippa Foot's recent argument in favour of a new approach to moral objectivity in virtue ethics, which is based on her concept of "natural goodness". If Foot's theory is philosophically defensible, it will allow the account of the categoricity of moral reasons given in the previous chapter to be supplemented with a plausible reason to believe in the existence of moral properties. Both elements of Joyce's sceptical challenge will thereby have been met, and moral realism will live to fight another day.

In §2 I provide some background details regarding the history of Foot's thought, and explain why her discussion of natural goodness constitutes a good starting point when looking for a realist reply to Joyce. As will be seen, Foot formerly, albeit briefly, occupied a philosophical position quite similar to the one endorsed by Joyce. Given that Foot was persuaded to abandon that position for one which she takes to be more objective in character, her arguments in defence of this change of position are worth paying serious attention to.

The exposition of Foot's theory of natural goodness begins in §3, which shows how Foot argues that it is necessarily practically rational to act in accordance with the virtues. In §4 I provide a detailed account of Foot's conception of natural goodness. Natural goodness is distinguished from secondary goodness, and I explain the way in which the concepts of natural goodness, Aristotelian necessities, Aristotelian categoricals, and flourishing are related to one another in Foot's thought.

§5 considers a number of objections to Foot's theory. I will show that the weakest links in Foot's theory are to be found in her welfarist conception of function, and in her commitment to the claim that it is possible to identify a universal form of human flourishing. I also argue that Foot's theory fails to establish a necessary connection between action in accordance with Aristotelian categoricals and the attainment of *eudaimonia*. The arguments surveyed in this section show Foot's theory to be indefensible.

Notwithstanding the failure of Foot's account, it might be thought that an alternative approach to virtue ethics is more likely to succeed in defending moral realism against Joyce. In §6 I consider just such an approach, as argued for by Jonathan Haidt and Craig Joseph. This account proves to be a philosophical improvement on Foot's theory, avoiding as it does the problematic issues of *eudaimonia* and non-Darwinian conceptions of function. However, Haidt and Joseph's theory is not proposed as a normative theory, but as a piece of descriptive moral psychology. It is therefore not well suited to the task of providing an alternative defence of moral realism. If such a defence is to be made, it therefore seems that it must be made outside of virtue ethics.

2. A New Route to Objectivity?

In *Natural Goodness*, Philippa Foot presents an account of the objectivity of morality, and of the status of moral reasons as categorically binding. Her account is an intuitively attractive starting point from which to look for a reply to Richard Joyce's sceptical challenge. There are two main reasons for this attractiveness.

Firstly, prior to her endorsement of an objective, categorical theory of morality, Foot subscribed to a view that has much in common with that of Joyce. In her widely read 1972 article, "Morality as a System of Hypothetical Imperatives," Foot argued that moral demands apply to an agent only insofar as that agent is committed to acting from within the institution of morality. Moral imperatives, on this view, are akin to the imperatives created by social institutions and conventions, such as the norms of etiquette. Yet such norms do not bind categorically. For example, norms of etiquette specify which item of cutlery ought to be used during the consumption of specific dishes. If, however, one is at an informal occasion, such as a friend's barbecue, the norms of etiquette may simply be set aside. In this instance, their strict enforcement would likely be an obstacle to the success of the barbecue: they would get in the way of everyone's having a good time. Thus, for Foot, norms of are etiquette are hypothetical imperatives. That is, their authoritative force is generated either by the desires and commitments of individual agents, or by those "belonging to a number of people engaged in some common project or sharing common aims".[1] The same is true, Foot holds, of moral norms:

> moral judgments have no better claim to be categorical imperatives than do statements about matters of etiquette. People may indeed follow either morality or etiquette without asking why they should do so, but equally well they may not. They may ask for reasons and reasonably refuse to follow either if reasons are not to be found.[2]

Thus Foot denies the claim that, to the extent to which she acts immorally, an agent must be considered irrational.

There are some terminological differences which, unless explicitly noted, might obscure the extent to which Joyce agrees with Foot's early work. Thus, whilst Foot characterises norms of etiquette and morality as hypothetical imperatives, Joyce refers to them as "weak" or "institutional" categorical imperatives. He does so in order to draw attention to the fact that such norms continue to apply to us, even in the event that we are not rationally obliged to adhere to them. For example, using a fork with one's right hand at a friend's barbecue still contravenes one of the rules of etiquette, even if nobody present at the barbecue has a reason to observe or enforce those rules. Weak categorical imperatives contrast with hypothetical imperatives in that the latter recommend a course of action as a means to an end: "if you want to get the next train, you had better leave for the station now". However, if it is not the case that you want to get the next train, then it is not the case that you ought to leave now. Here, the hypothetical imperative altogether ceases to apply; as Joyce puts it, "the imperative is retracted".[3] Joyce contrasts weak categorical imperatives with "strong" or "non-institutional" categorical imperatives. These are the imperatives which Foot simply describes

[1] Foot (2002a) p. 159.
[2] Foot (2002a) p. 164.
[3] Joyce (2001) p. 35.

as "categorical". That is, they are imperatives which "imply that persons have *reasons* to act regardless of their desires or interests [...] [they] *bind* persons, in a way that etiquette does not".[4]

As seen in the previous chapter, Joyce does not object to weak categorical imperatives; the target of his sceptical attack is the conceptual coherence of strong categorical imperatives. Despite the differences in their terminology, then, Joyce is largely in agreement with Foot *circa* 1972.[5] Yet in her later work, Foot rejects the argument of "Morality as a System of Hypothetical Imperatives". As she describes it, this position was "a bad mistake",[6] and was the product of:

> a despairing mood, [in which] I was even ready to deny that for everyone, always, it *would* be rational to act morally. This was often identified as "Foot's position", long after I myself had abandoned it and was working my way around, very slowly, to the quite different position that I first took up in the eighties and have held ever since.[7]

According to Foot, the position which she espouses in her later work allows her "to speak more robustly about objectivity"[8] than was previously the case. If Foot is correct in this estimation of her earlier work, then her arguments for a more robust conception of moral objectivity might also constitute a realist reply to Joyce.

The second reason to look for a reply to Joyce in Foot's later work is that it is explicitly evolutionary. This is a result not only of Foot's philosophical commitment to virtue ethics, but also of the historical origins of virtue ethics in Aristotle's naturalistic conception of the good human life.

According to Aristotle, the virtuous agent is the one whose actions occupy intermediate positions between opposing vices of excess and deficiency (this is sometimes referred to as the doctrine of the "golden mean"). For example, Aristotle holds that:

> [w]ith regard to pleasures and pains [...] the intermediate state is moderation, the excessive state self-indulgence. As for people deficient with regard to pleasures, they hardly occur; which is why [...] [they have] failed to acquire a name. But let us put them down as "insensate". With regard to the giving and receiving of money the intermediate state is open-handedness, while the excessive and deficient states are wastefulness and avariciousness.[9]

As a result of developing a reliable disposition towards virtuous agency, Aristotle argues, an agent will be more likely to achieve *eudaimonia*. This is a state of maximal happiness, the attainment of which is, for Aristotle, the *telos* of all rational action.

> Happiness [...] we do always choose because of itself and never because of something else, while as for honour, and pleasure, and intelligence, and every excellence, we do

---

[4] Joyce (2001) p. 37. Original emphasis.
[5] Though they disagree about the philosophical commitments built into ordinary moral discourse. For a more detailed discussion of this disagreement, and of the distinction between weak and strong categorical imperatives, see Joyce (2001) Chapter 2.
[6] Foot (2002b) p. 2.
[7] Foot (2002c) p. x.
[8] Foot (2002b) p. 2.
[9] Aristotle (2002) 1107b4 – 1107b11.

> choose them because of themselves [...] but we also choose them for the sake of happiness, supposing that we shall be happy through them.[10]

It is important to note that the *eudaimon* individual will not only possess internal, or moral, well-being. A certain amount of material comfort and physical health are also prerequisite conditions of *eudaimonia*. For this reason, *eudaimonia* is often translated as "flourishing" rather than as "happiness". As David Bostock explains, "one cannot be *eudaimon* without some wealth, nor without good looks, good birth, friends, political influence, and successful children".[11] Bostock describes these as "necessary conditions"[12] for the achievement of *eudaimonia*; it is worth pointing out, however, that it does not follow from their being necessary for *eudaimonia* that these external goods are themselves partly *constitutive* of *eudaimonia*. This is not an issue which I will explore here, however.

Modern, neo-Aristotelian treatments of virtue ethics typically do not characterise the virtues as intermediates between opposing vices. They do, however, follow Aristotle in claiming that possession of the virtues, and action in accordance with them, leads to *eudaimonia*. As will be seen, Foot attempts to cash out the notion of *eudaimonia* in terms of the proper functioning of our evolved capacities. This move has some *prima facie* appeal as a possible reply to Joyce. If the virtues are those character traits whose possession leads to human flourishing, then they would seem to be objectively grounded in biological data. That is, facts about human evolution would be *productive* of facts about human morality, insofar as it is the evolutionary process which determines *what it is* to be a flourishing human organism. Joyce's argument that evolution calls the existence of moral facts into question would therefore be shown to rest on a mistaken assumption; i.e. that such facts ante-date human existence. As I will show, however, Foot's argument is ultimately unsuccessful.

3. The Practical Rationality of Virtue

According to the Foot of *Natural Goodness,* and *contra* her earlier work, "acting morally is part of practical rationality".[13] Foot makes this claim in order to accommodate the motivation-internalist's intuition that there is a necessary connection between moral belief and moral motivation.[14] Foot claims that her previous reluctance to ground morality in practical rationality stemmed from her acceptance of "a more or less Humean theory of reasons for action"; i.e. from a perspective similar to the reasons internalism endorsed by Joyce and Bernard Williams.[15]

The account of practical rationality which Foot provides differs considerably from the one typically employed in philosophical discourse, however. Thus, Foot rejects those conceptions of the practical rationality of morality which are based on "claims of self-interest or maximum satisfaction

---

[10] Aristotle (2002) 1097b1 – 1097b5.
[11] Bostock (2000) p. 12.
[12] Bostock (2000) p. 12.
[13] Foot (2001) p. 9.
[14] Discussed in Chapter Three, §7 of this thesis.
[15] For discussion of which, see Chapter Three §6.ii of this thesis.

of desires [...] [and which see] the rationality of moral action in terms of the one that wins out".[16] Instead of this approach, Foot argues that it is possible to place conceptual constraints on the notion of practical rationality. With these constraints in place, the satisfaction of immoral desires can be automatically excluded as a candidate for potentially practically rational action. Unless the satisfaction of evil desires *can* be so excluded, Foot argues, it becomes impossible to identify what "would be *so important* about practical rationality".[17] According to the alternate conception of practical rationality which Foot argues for, then, "there is no criterion for practical rationality that is not *derived from* that of goodness of the will".[18]

Foot provides some support for her argument that immoral actions are necessarily practically irrational. She does this by claiming that it is one of the conceptual properties of a virtue that, "in so far as someone possesses it, his actions are good; which is to say that he acts well".[19] This is not to say, however, that it is conceptually impossible for a virtuous agent to *fail* to act well; it is simply the case that such failure will be a result of contingent features of the circumstances in which the virtuous agent finds herself. For example, she may be rendered incapable of fulfilling an honestly made promise because of the onset of a sudden illness, or as a result of being forcibly restrained. What identifies a virtuous agent as such is not, therefore, their unfailing performance of virtuous actions. Rather, it is that they identify certain features of a situation as providing them with decisive *reasons* for action. Thus, the virtuous agent will recognise that the making of an honest promise gives her a reason to keep that promise. For Foot:

> it is the distinguishing characteristic of the [virtuous] that *for them certain reasons count as reasons for action, and as reasons of a given weight* [...]. Those who possess [...] virtues possess them in so far as they recognize certain considerations (such as the fact of a promise, or of a neighbour's need) as powerful, and in many circumstances compelling, reasons for acting. They recognize the reasons, and act on them.[20]

So construed, Foot argues, the concept of a virtue *just is* part of the concept of practical rationality:

> The discussion [of the virtues] has been about human goodness in respect of reason-recognition and reason-following, and if this is not practical rationality I should like to know what is.[21]

As noted above, however, the virtues form only a *part* of Foot's conception of practical rationality. Its other aspects are supplied by means-end reasoning about desire satisfaction and the pursuit of self-interested goals; i.e. by the alternate accounts of practical rationality which are *prima facie* opposed to Foot's view. Foot claims that there is in fact no need to view these as *alternatives* to her approach. Rather, they may be seen as complimentary ways of filling out a single, inclusive account of what it is to be practically rational.

---

[16] Foot (2001) p. 11.
[17] Foot (2001) p. 10. Original emphasis.
[18] Foot (2001) p. 11.
[19] Foot (2001) p. 12.
[20] Foot (2001) p. 12. Original emphasis.
[21] Foot (2001) p. 13.

> An action can be contrary to practical rationality in that it is dishonest or disrespectful of others' rights, *or* that it is foolishly imprudent; *or*, again, that the agent is, for example, careless, timid, or half-hearted in going for what he wants.[22]

Foot's account presents practical rationality as a multi-faceted phenomenon. As a result, her account raises an immediate and important question: is it necessary that the different aspects of practical rationality are always concurrently realisable? This would certainly appear not to be the case. Suppose, for example, that Jack owes Jill some money. Jack has promised to repay his debt to Jill today, and they have arranged to meet for lunch so that he may do so. Now, further suppose that Jack is low on cash at the moment. He can still afford to repay Jill today, but this will mean that he must live frugally for the rest of the month. Jack would much prefer to repay Jill in two weeks' time, when he will have more money with which to do so. Jack also knows that Jill considers him to be somewhat profligate. He knows that she will not be unmoved by any request to defer repayment of his debt to her, even though Jill has no great need of the money at present; indeed, she might even see Jack's constrained financial situation as an opportunity for him to learn how to manage his money more efficiently. Jack enjoys spending money as he does, however, and has no desire to live what he sees as an austere existence. Having taken all this into consideration, Jack sends Jill an email cancelling their lunch-date, on the fabricated pretext that he is feeling unwell.

In the scenario sketched above, there appears to be a conflict between the different aspects of Jack's deliberation over what it is practically rational for him to do. On the one hand, in breaking his promise to Jill he fails to act virtuously, and is practically irrational on that account. Yet on the other hand, by acting virtuously Jack fails to act on a means to one of his ends; i.e. that of avoiding having to live frugally. But this, too, is a feature of practical rationality. Whatever he does then, it seems that Jack is fated to act in a practically irrational way. Whilst this might be a plausible conclusion in certain extreme circumstances,[23] it certainly does not seem reasonable in this scenario.

Clearly, what Foot needs in order to avoid this problem is a way of supplementing her account of the practical rationality of virtuous agency, such that it becomes unequivocally practically irrational to act on desires contrary to the virtues. Foot characterises the challenge as one of legitimately denying "that there is an independent criterion of rational action – the pursuit of happiness – with rationality on occasion demanding what virtue forbids".[24]

Foot's response to this challenge is to claim that "happiness can be seen [...] as conceptually inseparable from virtue".[25] Her defence of this claim is based on the conceptual connection between the term "happiness" and Aristotle's analysis of *eudaimonia*. According to that account, Foot notes, action in accordance with virtue "was the essence of the concept of *eudaimonia*".[26] However, it does not follow from this, Foot claims, that the virtuous agent will necessarily be a happy one. This is because the conditions under which an agent exercises her virtuous dispositions may preclude her from living a happy life. This, Foot argues, was the case for those individuals unfortunate enough to find themselves living in Nazi occupied territories during the Second World War. Such individuals

---

[22] Foot (2001) p. 13.
[23] I have in mind here scenarios of the kind that Hursthouse terms "tragic dilemmas", in which none of the actions available to an agent are morally permissible. See Hursthouse (1999) Chapter 3 for discussion.
[24] Foot (2001) p. 82.
[25] Foot (2001) p. 94.
[26] Foot (2001) p. 97.

could, when called upon to do so, have agreed to aid the Nazis. Refusal to do so would have brought with it the dangers of imprisonment, torture, and death. Yet complicity with the Nazis would, for these virtuous agents, have precluded the possibility of happiness, simply because it would have compromised their virtue. Thus Foot holds that "humanity's good can be thought of as happiness, and yet in such a way that combining it with wickedness is *a priori* ruled out".[27] That is, *eudaimonia* is impossible without possession of the virtues. Yet, given Foot's argument that there is a "conceptual connection between acting well and acting rationally",[28] it is precisely this at which all of our rational actions aim. Accordingly, when our desires conflict with virtuous agency, it is irrational to satisfy them.

So far, we have seen how Foot argues that virtuous action is practically rational action, and that cultivation of the virtues is necessary for the attainment of *eudaimonia*. That is to say, Foot identifies the virtues as those dispositions which are necessary for human flourishing. She claims that, for any given member of a species:

> [t]o flourish is [...] [for that individual] to instantiate the life form of that species, and to know whether an individual is or is not as it should be, one must know the life form of the species. A quite general conceptual connection between life form and goodness is given specification in the myriad life forms of different kinds of living things; no doubt historically by an evolutionary story that leaves the members of each species dependent not only on their own internal resources but also on the environment to which the species came to be adapted.[29]

It is this view which underlies Foot's account of how virtuous dispositions are to be identified as such. Through their contribution to a specifically human way of life, the virtues contribute to their possessor's "natural goodness";[30] i.e. to their flourishing. The virtues' connection to natural goodness allows Foot to claim that they can be identified as genuinely objective: the virtues are reified by facts about human biology, and the requirements of human societies. To see more clearly how Foot argues for this claim, more must be said about her conception of natural goodness, and of what it consists in. This will be the task of the next section.

4. The Nature of the Virtues

Foot describes her principal aim in *Natural Goodness* as being to "describe a particular type of evaluation and to argue that moral evaluation of human action is of this logical type".[31] It is this type of evaluation which we use to assess a living organism in terms of its "natural goodness". What, then, *is* natural goodness, and how does it differ from other ways of evaluating living organisms?

---

[27] Foot (2001) p. 96. Italics added.
[28] Foot (2001) p. 65.
[29] Foot (2001) pp. 91 – 92.
[30] Foot (2001) p. 3.
[31] Foot (2001) p. 3.

(i) Natural and Secondary Goodness

Foot answers this question by drawing a distinction between natural and "secondary" goodness, the latter type of goodness being that which she takes her readers to be more familiar with. Foot describes secondary goodness as "derivative"[32] goodness, though, given the examples which she uses to illustrate this concept, the term "relational" might be more appropriate. Secondary goodness, then, is the goodness which is predicated of a thing (and not necessarily a *living* thing) when it serves some purpose which is beneficially related to the activity of some organism. As this is quite an abstract formulation, some examples will help to illustrate the basic idea.

We might say of a certain soil type that it is good for growing crops such as maize; similarly we might say that dry weather conditions are good for the planting of maize, or that maize is good for feeding livestock. All these predications of goodness pertain to *secondary* goodness. The thing which is described as good, be it soil, the weather, or maize, is so described only to the extent that it promotes a specified end. Secondary goodness is therefore contingent goodness. If it was not the case that maize was good for feeding livestock, then it would not possess secondary goodness in that respect. Furthermore, secondary goodness can be lost without the thing of which it is predicated itself undergoing any change. To see this more clearly, consider penicillin. When it was first discovered, penicillin was a highly effective antibiotic. It therefore had a high degree of secondary goodness. As penicillin has become more widely used, however, many bacteria have evolved to become resistant to it. If this trend continues, penicillin will ultimately become ineffective against many common infections. At this point, it will no longer possess secondary goodness. Yet penicillin *itself* will be the same as it ever was. Rather, the environment which formerly conferred secondary goodness upon it will have changed.

Natural goodness, on the other hand, is "intrinsic or 'autonomous' goodness";[33] it "depends directly on the relation of an individual to the 'life form' of its species".[34] That is, the possession of natural goodness has "nothing to do with the needs or wants of the members of any other species of living thing".[35] When we judge that an individual member of some species possesses the property of natural goodness, Foot claims, we do so by drawing on a plethora of connected pieces of information. The aggregate of this information constitutes our conception of *how members of that species ought to be*. That is:

> *[g]oodness* in plants and animals nests in an interlocking set of general concepts such as *species, life, death, reproduction,* and *nourishment*, together with less general – we might say local – ideas such as that of *fruiting, eating,* or *fleeing*.[36]

---

[32] Foot (2001) p. 26.
[33] Foot (2001) p. 27.
[34] Foot (2001) p. 27.
[35] Foot (2001) p. 26.
[36] Foot (2001) p. 36. Original emphasis.

To help make this account more precise it is necessary to introduce two additional concepts, both of which are central to Foot's conception of natural goodness. These are "Aristotelian necessities" and "Aristotelian categoricals".

(ii) Aristotelian Necessities

Foot, drawing on the work of G. E. M. Anscombe,[37] defines an Aristotelian necessity as "that which is necessary because and in so far as good hangs on it".[38] Aristotelian necessities describe the species-typical requirements upon which an organism depends if it is to engage in the activities which are characteristic of members of its species. This is a highly abstract formulation, but such abstraction is unavoidable when the concept being elucidated encompasses such a diverse array of phenomena. Some concrete examples of particular Aristotelian necessities are therefore in order. Foot provides the following, claiming that:

> we invoke the same idea [i.e. of an Aristotelian necessity] when we say that it is necessary for plants to have water, for birds to build nests, for wolves to hunt in packs, and for lionesses to teach their cubs to kill.[39]

It is worth highlighting at this point that the species-specific nature of Aristotelian necessities means that what counts as such for one species need not do so for another. Not all birds make nests, for example, and so (*contra* the implication of Foot's example) nest-making is not an Aristotelian necessity for every species of bird.

Aristotelian necessities are determined by "what the particular species of plants and animals need, on their natural habitat, and on the ways of making out that are in their repertoire".[40] Once these necessities have been identified, Foot claims, it is possible for an observer to "determine what it is for members of a particular species to be as they should be, and to do what they should do".[41] If an organism satisfies these criteria, then it possesses natural goodness. Conversely, any organism which is *not* as members of its species ought to be can be considered "defective": "free-riding individuals of a species whose members work together are just as *defective* as those who have defective hearing, sight, or powers of locomotion".[42] This type of evaluation, Foot claims, holds good for any living organism, and human beings are no exception: "I am therefore, quite seriously, likening the basis of moral evaluation to that of the evaluation of behaviour in animals."[43] It is, Foot argues, essential that human beings live in accordance with the virtues.

---

[37] See Anscombe (1981a; 1981b)
[38] Foot (2001) p. 15.
[39] Foot (2001) p. 15.
[40] Foot (2001) p. 15.
[41] Foot (2001) p. 15.
[42] Foot (2001) p. 16. Original emphasis.
[43] Foot (2001) p. 16.

> Anyone who thinks about it can see that for human beings the teaching and following of morality is something necessary. We can't get on without it.[44]

For humans, that is, the cultivation of virtue is an Aristotelian necessity.

Yet there is more to Foot's account of normative evaluation than the identification of Aristotelian necessities. As noted, these describe the *needs* of an organism: the requirements which must be met for it to flourish. Yet failure to obtain the Aristotelian necessities which one needs in order to flourish need not be morally culpable: witness Foot's claim that *eudaimonia* might be impossible to attain in Nazi Germany, even if one possesses virtuous dispositions (see §3). To make room for culpability in her account, Foot introduces the concept of *Aristotelian categoricals*.

(iii) Aristotelian Categoricals

Whilst Aristotelian necessities pertain to an organism's *requirements*, Aristotelian categoricals pertain to its physical and behavioural *traits*. Aristotelian categoricals are propositions which describe species-typical traits:

> [t]hey tell how a kind of plant or animal, considered at a particular time and in its natural habitat, develops, sustains itself, defends itself, and reproduces.[45]

Yet according to Foot's account, not all species-typical traits are to be described using Aristotelian categoricals.[46] Only propositions pertaining to traits which "play a part"[47] in the life of the species are to be thought of as Aristotelian categoricals. For a trait to "play a part" in the life of a species is, according to Foot, for it to be either "directly or indirectly [...] causally and teleologically related"[48] to the survival and reproduction of the members of that species. Accordingly, the proposition "peacocks have brightly coloured tails" is an Aristotelian categorical, whereas the proposition "peacocks have grey feet" is not.[49]

Foot argues that their relation to the survival and reproduction of a species means that Aristotelian categoricals "describe norms rather than statistical normalities".[50] Despite their close connection to these ends, failure to act in accordance with Aristotelian categoricals need not always have a negative consequence for an organism. However, even if this is the case, that organism is still to be considered a defective token of its type.

---

[44] Foot (2001) pp. 16 – 17.
[45] Foot (2001) p. 29.
[46] In this respect, Foot's position differs from that of Michael Thompson. Otherwise, her account is largely derived from Thompsons'. For the latter, see Thompson (1995).
[47] Foot (2001) p. 31.
[48] Foot (2001) p. 31.
[49] On the assumption, that is, that foot colouration does not relate to a peacock's survival or reproductive success.
[50] Foot (2001) p. 33.

> Take, for instance, the dance of the honey bee which tells other bees of a source of food. No doubt an individual bee that does not dance does not itself suffer from its delinquency, but *ipso facto* because it does not dance, there is something wrong with it, because of the part that dancing plays in the life of this species of bee.[51]

Foot's reasoning can be summarised as follows: Dancing plays a part in the life of the honey bee species, by allowing members of that species to procure some of their Aristotelian necessities. Specifically, dancing allows honey bees to indicate sources of food to one another. It is therefore an Aristotelian categorical that "honey bees dance to indicate the location of food". Thus, for a honey bee to possess natural goodness it must be one which (among other things) performs such a dance. Honey bees *ought* therefore to dance; not to do so is to be a defective, i.e. non-flourishing, honey bee.

It is important to note the shift in focus that occurs when talking of Aristotelian categoricals, as opposed to Aristotelian necessities. The latter specify what an organism *needs*, whereas the former specify how an organism ought to *act*. Normative judgements about plant development and animal behaviour, therefore, are to be based on our conception of the Aristotelian categoricals which describe how such organisms ought to be. The normative evaluation of human action, Foot contends, is in principle no different from this way of evaluating non-human life-forms. Just as we describe plants and animals as good or bad to the extent that they possess natural goodness, so too our moral evaluation of human agency is "of this logical type".[52] Foot concedes that such a view is likely to "provoke instant opposition".[53] Before turning to a critical analysis of her position, then, something ought to be said regarding Foot's argument for applying the concept of natural goodness to the moral evaluation of human actions.

(iv) Natural Goodness and Human Morality

Foot does not give a detailed account of which dispositions her theory identifies as virtuous. Rather, she limits her task to a defence of the legitimacy of evaluating human actions in terms of their natural goodness.

Foot does not claim that natural goodness in humans pertains solely to the capacity to survive and reproduce. She thus argues that:

> choice of childlessness and even celibacy is not thereby shown to be a defective choice, because human good is not the same as plant or animal good.[54]

In fact, Foot is reluctant to specify too precisely what the human good *does* actually consist in, remarking that the issue is one which is "deeply problematic".[55] The variety of habitats which

---

[51] Foot (2001) p. 35.
[52] Foot (2001) p. 3.
[53] Foot (2001) p. 38.
[54] Foot (2001) p. 42.
[55] Foot (2001) p. 43.

humans inhabit, and the cultural diversity manifested across the range of known human societies, makes providing a single, definitive account of human flourishing a daunting task. This is not a task which Foot feels the need to undertake, however. Whatever the nature of the human society under consideration, she holds that its members will need to possess certain characteristics if we are to describe them as flourishing. These characteristics include, but are not limited to, the virtues. For example, Foot claims that:

> human beings need the mental capacity for language; they also need powers of imagination that allow them to understand stories, to join in songs and dances – and to laugh at jokes. Without such things human beings may survive and reproduce themselves, but they are deprived. And what could be more natural than to say on this account that we have introduced the subject of possible human defects; calling them "natural defects" as we used these terms in the discussion of plant and animal life?[56]

Foot then goes on to argue that the virtues are no less essential for a flourishing human life than is the capacity for language:

> Men and women need to be industrious and tenacious of purpose not only so as to be able to house, clothe, and feed themselves, but also to pursue human ends having to do with love and friendship. They need the ability to form family ties, friendships, and special relations with neighbours. [...] And how could they have all these things without virtues such as loyalty, fairness, kindness, and in certain circumstances obedience?[57]

Foot's argument, it will be recalled, holds that virtuous dispositions are necessary preconditions for the attainment of *eudaimonia*, and that the attainment of *eudaimonia* is that at which practically rational action aims. Accordingly, when an agent recognises that a particular virtue is necessary for the possession of natural goodness, and that it is in terms of their natural goodness that we identify organisms as flourishing, or *eudaimon*, the agent in question then has a reason to act in accordance with that virtue. Failure to so act, either as a result of weakness of will or of a constitutive inability, identifies that agent as rationally defective, and therefore as a legitimate target of moral censure.

Foot's analysis of the virtues, if correct, would constitute a reply to Joyce's sceptical challenge. Joyce uses evolutionary data to cast doubt upon the existence of objective moral properties. According to Foot, however, human evolution is itself responsible for the existence of objective moral properties. Human evolution, both biological and cultural, determines what it is for humans to live flourishing lives. This is simply a fact about human nature. Other facts, facts about the sorts of character traits necessary for human flourishing to obtain, determine which of those character traits are to be identified as virtuous. As all our rational actions aim at our flourishing, so too must they all aim at virtue. Any agent who recognises that a particular action is virtuous, also recognises that she has a reason to perform that action. If she fails to recognise this then, as Foot expresses it elsewhere, she "fails to recognize the truth".[58] She is rationally defective.

---

[56] Foot (2001) p. 43.
[57] Foot (2001) pp. 44 – 45.
[58] Foot (2002d) p. 170.

With this outline of Foot's account in place, the next section considers some of the objections which can be made in response to it. As these will show, Foot's account is in fact deeply problematic. Thus, whilst Foot's theory of natural goodness does constitute a possible reply to Joyce, nevertheless it is not a theory which ought to be endorsed.

5. Objections to Foot's Theory

In this section I consider four objections to Foot's account. The first two of these objections are made by Joyce. Joyce's first objection is based on the claim that virtue ethics cannot give a satisfactory account of moral culpability. This objection is not successful, but it is worth discussing as doing so helps to avoid a potential source of confusion regarding what, according to Foot, makes an action immoral. The second of Joyce's objections is more compelling, and identifies a serious issue with the coherence of virtue ethics' attempt to identify a universal form of human flourishing. The third objection which I consider, made by Tim Lewens, also poses a significant challenge for Foot's theory. Lewens argues that Foot's welfarist account of function is unable always to identify the causes of flourishing, and that it is therefore unable to identify virtuous agency. The fourth and final objection to Foot's account which I consider is one which I develop from Lewens' criticism of Foot's welfarist notion of function. There, I argue that Foot's claim that an organism must act in accordance with its Aristotelian categoricals in order to flourish is empirically false. Given the centrality of this claim to Foot's account of the virtues, its falsity shows her theory of natural goodness to be unworkable.

(i) The Moral Culpability of Vicious Agency

Both of Joyce's objections are objections to virtue ethics *tout court*. The first of these objections is that virtue ethics is unable to provide a satisfactory account of moral culpability, and that this shows the virtue ethicist to be working with an idiosyncratic conception of morality. Specifically, Joyce argues that virtue ethics portrays moral culpability as arising from the harm which a vicious agent inflicts upon *herself*. He argues that this is a consequence of virtue ethics' claim that an agent acts morally in order to attain *eudaimonia*.

Virtue ethics, according to Joyce, consists of a series of hypothetical imperatives, albeit ones which are "in some sense inescapable".[59] These imperatives are inescapable because every agent has the *de facto* end of attaining *eudaimonia*. Yet they are hypothetical imperatives because their validity is nevertheless contingent upon an agent having that end. If, *per impossibile*, there were an agent who did not wish to flourish, then she would not be practically irrational to disregard the reasons for action which the virtues provide.

[59] Joyce (2006) p. 171.

> The basic problem is that someone who fails to act so as to secure her ends has principally wronged *herself*, but a value system revolving around self-harm doesn't look much like a moral system. Punishment, in such a system, would amount to a bizarre institution of inflicting further harm upon a person because she has harmed herself. On such a view Jack the Ripper should elicit our deepest sympathy, since what was really wrong about his killing all those women is that he radically undermined his own flourishing. ("Poor man," we should say upon capturing him.)[60]

Joyce's objection fails however, at least when applied specifically to Foot's brand of virtue ethics. This is because Joyce illegitimately conflates an agent's *reason* for acting morally with why it is *morally wrong* to act *im*morally. It was noted earlier (in §3) that, according to Foot, the concept of practical rationality is not *wholly* a matter of recognising the reasons given by the virtues. One can act in a practically irrational way by (amongst other things) being imprudent, being reckless, or by being insufficiently zealous in striving after one's goals. Yet these failures of practical rationality are not obviously *moral* failures. This status is reserved for those actions which are not *only* practically irrational, but which are *also* harmful to others. It is the harm caused by such actions which identifies their perpetrators as morally culpable. The practical irrationality of immoral actions is *not* what makes them immoral; rather, their irrationality is what gives an agent a reason not to commit them. Of course, Joyce will still wish to press his point that virtue ethics gives rise only to *hypothetical* imperatives, however inescapable these may appear to be, and that morality proper deals with *categorical* imperatives. I have argued against Joyce's conception of categorical imperatives in the previous chapter, and will not rehearse that argument here. Suffice it to say, I think that when that argument is taken into consideration, the first of Joyce's objections to virtue ethics loses its force.

(ii) The Hidden Error in Virtue Ethics

The second of Joyce's objections fares somewhat better than the first. In it, he argues that virtue ethics rests on a conceptual mistake: the virtue ethicist is committed to the existence of virtuous character traits, but in fact no such traits, at least as conceived of by the virtue ethicist, exist. It should be noted that in making this objection Joyce does *not* endorse the view that there are no stable character traits; i.e. that our actions are wholly determined by the circumstances that we happen to find ourselves in, and not by stably recurring patterns of psychological reaction. Rather, Joyce's argument is based on a problem which he takes to be generated by virtue ethics' commitment to *eudaimonism*.

Virtue ethics, Joyce writes, identifies the virtues by asking what kind of life an agent ought to live, or what kind of person an agent should be. The virtues are then identified as such by the fact that they "have the theoretical role of *answers* to these questions".[61] Yet, Joyce argues, such questions do not in fact admit of definite answers in the way supposed by the virtue ethicist. This is

---

[60] Joyce (2006) p. 171.
[61] Joyce (2011) p. 173.

because the range of human cultures, and the various conceptions of "the good life" that go with them, are vast. It is unlikely, Joyce contends, that the numerous lists of virtuous dispositions generated by each of these rival conceptions will be isomorphic with one another. Yet if they are not, then there is no answer to the question to which virtue ethics is a response.

Joyce explicates his position by way of reference to a slightly whimsical analogy. Thus he asks:

> *What kind of ice cream flavor must one prefer in order to be the kind of person one should be? What ice cream preference contributes to the good life for a human being?* Let us assume, not unreasonably, that it is acceptable to choose ice cream flavors on the basis of gustatory whim. [...] Then I would know what flavor *I* should prefer, but there would be no flavor that *one* should prefer, and no flavor that one must prefer in order to be the kind of person that one should be.[62]

On any plausible account of the virtues, Joyce argues, virtuous dispositions "cannot be similarly a matter of whimsical choice and cannot change from individual to individual",[63] yet this is precisely what *does* happen given the range of recognisably good human lives. Thus, the virtues needed to live the life of an ascetic saint will not be the same as those needed to live the life of a politician, and neither set of these will be the same as those needed to live the life of a member of a self-sufficient commune of eco-warriors.

Joyce notes that the virtue ethicist will wish, at this point, to insist that all of these ways of life require the development of a core set of virtues, which can be considered essential to *any* good human life. Joyce doubts that such a move can be successful, however. He argues that human psychological development is so open-ended that the concept of *eudaimonia* "may provide only a minimal constraint on lifestyle decisions, and no constraint at all on character traits".[64] He supports this claim by arguing that social nature of human beings is in fact compatible with a wide range of vicious character traits.

> Hitler had loyal and sincere admirers; Genghis Khan was surrounded by good mates; perhaps even Jack the Ripper was a solid family man. The idea that the sociality inherent in human nature cannot be satisfied in a restricted domain, while coupled with cold disregard and astounding cruelty toward anyone lying outside the favored sphere, strikes me as a romantic misapprehension. [...] If this is so, then there may be no specific set of character traits that is underwritten by our social nature.[65]

If this is the case, then there can be no definitive answer to the question "which character traits ought I to possess?" Given, however, that it is a part of the concept of the virtues that they provide the answer to this question, it follows that there are no virtues. Thus, virtue ethics is committed to an error.

---

[62] Joyce (2011) p. 173.
[63] Joyce (2011) p. 173.
[64] Joyce (2011) p. 174.
[65] Joyce (2011) p. 174.

Joyce's objection poses a serious problem for Foot's position, particularly given her cautionary note that "the idea of a good life for a human being, and the question of its relation to happiness, is each deeply problematic".[66] Can Joyce's objection nevertheless somehow be met? The most likely way of doing so is to draw attention to the all-or-nothing aspect of *eudaimonia*. One cannot be *eudaimon* to a greater or lesser extent. As Bostock puts it, "[o]ne is either *eudaimon* or not, absolutely".[67] With this in mind, and given Foot's claim that "happiness can be seen [...] as conceptually inseparable from virtue",[68] it is possible to question whether agents such as Jack the Ripper or Genghis Khan are likely to be truly *eudaimon*. Do the (putatively) vicious character traits which these agents possess prohibit them from participation in some activities which would improve their quality of life? Do some of the psychological dispositions which Jack the Ripper lacks render him less able to appreciate certain works of art or literature? I will not attempt to answer these questions here. Even were I to try, however, it is by no means obvious that these questions can be answered in a sufficiently non-question-begging way. After all, is it not possible for Joyce simply to stipulate that his version of Jack the Ripper be just as moved by *Anna Karenina* as any of its other readers, notwithstanding Jack's murderous proclivities? Any response to Joyce of the sort I have just suggested is therefore likely to prove highly controversial.

The best way for the virtue ethicist to respond to Joyce's objection is simply to avoid it altogether. That is, the virtue ethicist ought to jettison the role played by *eudaimonia* in identifying the virtues. As Joyce himself observes, his objections to virtue ethics "do not apply to any such versions".[69] Given the role which *eudaimonia* plays in Foot's conception of virtuous action as practically rational action, however, this is not a move which is available to her. Joyce's objection therefore constitutes a serious problem for Foot's account.

(iii) Foot's Theory Cannot Reliably Identify the Traits Which Contribute to Eudaimonia

Suppose the second of Joyce's objections could somehow satisfactorily be met, and it could be shown that it does in fact make sense to speak of a sufficiently generalised notion of flourishing. Nevertheless, it may still be the case that our conception of flourishing cannot be made sufficiently precise for it to be put to the use which Foot intends. This point is made by Lewens, who argues that Foot's account is unable always to identify which of an organism's various characteristics contribute to its flourishing. To see the full force of this objection, a brief digression concerning Foot's use of the term "function" will be necessary.

Foot does not talk about the function of an organism's traits in a Darwinian way. According to Darwinian conceptions, functions come with historical baggage. Joseph Millum provides an

[66] Foot (2001) p. 43.
[67] Bostock (2000) p. 11.
[68] Foot (2001) p. 94.
[69] Joyce (2011) p. 178.

example of a Darwinian conception of function with reference to sparrow colouration.[70] As he explains:

> sparrows have mottled brown plumage because ancestral sparrows that had such plumage were selected over sparrows with different coloured feathers; they were selected for because the plumage provided camouflage and so reduced predation. Hence the colour pattern of the sparrow has the [...] function of camouflage.[71]

Millum emphasises that according to such a view:

> function is a historical concept, that is, in order for something to have a [...] function it must have a causal history that involves at least one of its ancestors performing the function.[72]

Yet history plays no part in Foot's definition of a function. Thus she writes that:

> [t]he history of a species is not, however, the subject with which Aristotelian categoricals deal. Their truth is truth about a species at a given historical time, and it is only the relative stability of at least the most general features of the different species of living things that makes these propositions possible at all.[73]

As John Hacker-Wright describes it, "the notion of function Foot puts forward is welfarist".[74] This is not to say, however, that all of the functions in which Foot is interested are those from which an organism *itself* derives some benefit: the functions in question may instead contribute to the welfare of other members of the species. Nevertheless, the notion of welfare, construed as the needs either of an organism or of the species to which it belongs, is central to Foot's analysis of function.

> Foot's conception of function [...] picks out features that are part of an organism as they function in its species-characteristic life. Our conception of species-characteristic life defines a conception of the needs of members of that species against which we assess individual organisms.[75]

Foot's conception of function therefore *needs* to be non-Darwinian if it is to support her discussion of natural goodness. Foot's welfarist notion of function grounds her account of Aristotelian necessities in a way that Darwinian notions of function are simply unable to do. What, however, does any of this have to do with Lewens' objection that Foot's account is unable to identify the flourishing agent?

---

[70] Millum's example relates, specifically, to Ruth Garrett Millikan's notion of a "proper function". I do not go into a discussion of the concept of proper functions here, as this would take me too far from the objection to Foot which this example is intended to clarify.
[71] Millum (2006) p. 203.
[72] Millum (2006) pp. 203 – 204.
[73] Foot (2001) p. 29.
[74] Hacker-Wright (2009) p. 312.
[75] Hacker-Wright (2002) p. 313.

Lewens introduces his objection by observing that individual organisms make trade-offs when choosing[76] between the life cycles available to them. Thus, Lewens asks whether such an organism should:

> live for a long time, thereby using resources that would have enabled it to have more offspring? Or should it have a small number of healthy offspring, and then die shortly afterwards, leaving resources to future generations? What determines whether reproduction or self-maintenance is the more important element of flourishing? Foot seems committed to the claim that natural facts alone can decide the issue.[77]

If Foot subscribed to a Darwinian conception of function, then the question posed by Lewens might prove answerable. For example, the relative fitness of each of the reproductive strategies which Lewens outlines could be calculated, with the fittest strategy being deemed optimal. Yet, as we have just seen, this is not an option for Foot. She cannot appeal to a Darwinian conception of function without giving up one of the cornerstones of her theory of natural goodness. Yet, from the welfarist conception of function to which Foot subscribes, there is no good way of identifying which of the reproductive strategies we have been considering actually leads to an organism's flourishing.

> Consider a species in which the members with the greatest reproductive output die earlier than they need to, but have more offspring in virtue of this. Other members of the species live considerably longer, but they have somewhat fewer offspring. It is implausible to think that these latter members fail to flourish simply because their reproductive output is lower than that of fitter conspecifics. [...] It is not clear that any naturalized theory of function can tell us, in the face of trade-offs such as these, where flourishing lies.[78]

Lewens' recognises that according to Foot, the flourishing of humans is much more complex than the flourishing of non-human organisms. It might therefore be thought that issues such as this do not arise when dealing with the human good, which is not restricted to self-maintenance and reproduction. Yet an appeal to this additional complexity is unable to meet Lewens' criticism.

> Even if the case for humans has many more dimensions than plant cases, it is important for Foot that her account should begin by making sense of the relatively simple evaluations we use when assessing parts of plants. But the problem raised by the case of trade-offs between self-maintenance and reproduction is that even in the simple cases of plants and animals, it seems that evaluations based on theories of biological functioning do not deliver plausible verdicts with respect to normative notions of the flourishing of a plant or animal.[79]

Indeed, the increasing complexity of the human good is, if anything, intuitively likely to *exacerbate* the problem which Lewens identifies.

---

[76] "Choosing" here should of course be understood in the biologist's usual metaphorical sense; i.e. not as a reasoned choice following rational deliberation, but in the sense of an organism's more-or-less instinctive "opting for" or "embarking upon" a particular strategy.

[77] Lewens (2010) pp. 469 – 470.

[78] Lewens (2010) p. 470.

[79] Lewens (2010) pp. 470 – 471.

This issue raises a serious difficulty for Foot's account. If her notion of function is unable always to identify the traits which contribute to flourishing, then it is also unable always to identify which traits are the subject of a corresponding Aristotelian categorical. As this is the means whereby the virtues are identified as such, Foot's theory is therefore unable to give a full account of the virtues.

It might be argued that this is not a fatal objection to the utility of Foot's position. For example, if Foot's account has shown virtuous agency to be a constitutive aspect of human flourishing, and has thereby also shown that it is always practically rational to act virtuously, perhaps her theory has done all that it needs to. That we may, in some extreme circumstances, be unable to determine the virtuous course of action is simply a result of our limited faculties, and is no more or less a problem for Foot's account than it is for any other normative ethical theory. This line of argument is unsuccessful, however. This is because Foot's account does not make determining the virtuous course of action a matter of reasoning correctly; rather, the virtues are to be determined by the observation of empirical facts about how the members of a species behave. The difficulty here is therefore not the same as the difficulty involved in, say, correctly identifying which action lies in accordance with Aristotle's golden mean. Lewens' objection cannot therefore be so easily dismissed: Foot's account seems unable *even in principle* to identify all of the virtues, and if this is the case her account fails.

A different way in which a supporter of Foot might reply to Lewens would be to claim that he has overstated the difficulties which her account faces. According to this line of reply, making accurate empirical observations is by no means always a straightforward affair. Even if it is sometimes extremely difficult to determine which alternative, when faced with a trade-off, will lead to flourishing, it by no means follows that such determination is *impossible*. As long as such empirical impossibility remains un-established, the reply continues, there is still hope for Foot's theory. A reply of this sort has something of an air of desperation about it. Rather than giving us a reason to be hopeful of success, it seems only to recommend determinedly applying Foot's theory to scenarios to which it seems fundamentally ill-suited. Still, the availability of such a reply makes it possible to insist that, at least without further empirical research, Lewens' objection is not decisive. Lewens' worries about Foot's conception of function and its relation to flourishing can be developed in a different direction, however. I give an example of how this can be done in the next, and final, objection to Foot's account. There, I show that her theory is unable to derive a satisfactory account of flourishing from the notion of Aristotelian categoricals. As this is the cornerstone of Foot's theory, such an objection is fatal objection to her position.

(iv) Aristotelian Categoricals Are Only Contingently Related to Flourishing

The a-historical nature of Foot's conception of function, already discussed in relation to Lewens' objection, creates a further, more serious problem when closer attention is paid to the notion of an Aristotelian categorical. Aristotelian categoricals, it will be recalled, are propositions which describe the characteristic life form of a species. According to Foot's theory of natural goodness, any individual organism which does not act in accordance with the Aristotelian

categoricals predicated of its species is incapable of flourishing, and must therefore be considered defective. This claim, as I will now show, is simply false. Acting in accordance with species-specific Aristotelian categoricals is not a prerequisite of flourishing, and organisms which do not so act cannot therefore be deemed defective.

The claim that the relation between acting in accordance with Aristotelian categoricals and flourishing is only a contingent one can most easily be shown by considering a hypothetical example of learned behavioural change. Once these examples have allowed this claim to be explicated in a quite general way, it can be shown to apply to numerous other, non-hypothetical scenarios. Suppose there to be, then, a frugivorous species of primate. These primates eat only a small range of fruits, and as a result there is always intense competition for food resources. It is therefore an Aristotelian categorical that "members of this species live exclusively on a diet of fruit". Further suppose, now, that a member of this species is born with the ability to digest nuts. This individual can therefore exploit an additional food supply found in its environment, and because there is no competition for nuts from its conspecifics, it eats mostly nuts and not fruit. This individual has more reliable access to food than do any of its conspecifics, and is therefore better adapted to its environment. We can further hypothesise that this individual goes on to leave a higher than average number of offspring as compared to the rest of its conspecifics, and that these offspring not only inherit the ability to eat nuts, but also a preference for doing so (perhaps via mechanisms of behavioural inheritance, discussed in Chapter One §5.iii). Clearly, it seems, this primate and its offspring are flourishing. Yet they are doing so not only *in spite* of violating one of their Aristotelian categoricals, but precisely *because* of so violating it. This point can be more succinctly expressed by saying that Foot's notion of an Aristotelian categorical is unable to accommodate the phenomenon of adaptive behavioural plasticity.

Another example of this phenomenon, similar to that of the hypothetical primate species just given, was in fact observed in recent history. Blue tits, whose typical food supply consists of spiders and insects, learned during the first half of the twentieth century to pierce the foil caps on milk bottles, thereby gaining access to the creamy fat which had risen to the top of the bottle. As Eva Jablonka and Marion Lamb explain, this behavioural change was not the result of a genetic adaptation, but was in fact a learned behavioural response. Furthermore, this learning was not imitative, as no single bottle-opening technique was transmitted throughout the tit population.

> What naïve birds learned through watching experienced ones was that milk bottles are a source of food. They learned *how* to open the bottles by individual trial and error, each developing its own, idiosyncratic technique.[80]

Yet it does not make sense to say that "tits will open milk bottles to consume cream" is a species-specific Aristotelian categorical. Were cream-consuming tits therefore defective? Surely not. Indeed, the identification of an additional food-source made it more likely that cream-consumers, rather than their naïve conspecifics, would flourish.

It might at this point be thought that I am relying too heavily on unduly specific Aristotelian categoricals. Perhaps the cream-consuming tits *were* acting in accordance with an Aristotelian categorical, if this were more generally construed as "will eat spiders and insects, or scavenge for

[80] Jablonka and Lamb (2005) p. 170. Emphasis added.

other food supplies". This suggestion might have some plausibility when considering the foregoing examples, but it does not withstand application to more radical instances of cultural evolution. Consider, for example, the discovery of cooking. This was an incredibly beneficial instance of cultural evolution, allowing for much easier digestion of food,[81] and the destruction of potentially harmful bacteria. Given the novelty of this cultural innovation, the first tribe of proto-humans to begin cooking their food must, according to Foot's theory, have been naturally defective. Yet quite clearly this would not have been an obstacle to their flourishing. Consider, too, the first community of early humans to form a small settlement, and take up agriculture. At that time, it would have been an Aristotelian categorical that "humans live in nomadic tribes of hunter-gatherers". Those of our ancestors who embarked upon an agricultural way of life were then naturally defective. And yet, as we are here to testify, they flourished.

I do not see a way to re-phrase the propositions expressed by Foot's Aristotelian categoricals so as to enable them to cover the sorts of examples I have just considered. Any attempt to do so would make the content of those propositions so vague as to be almost totally uninformative, and therefore useless as potential guides to action. Consider, for example, "humans consume food for nutrition" as a putative Aristotelian categorical. This certainly covers hunting and gathering, as well as cooking. It would not entail, therefore, that humans who cooked their food were naturally defective, and so unable to flourish. However, such a categorical is so vague that it does not do anything to distinguish the eating behaviours of humans from those of other animals. It does not describe a species-typical trait, in the way that "rabbits eat grass" does. In fact, it amounts to little more than a re-statement of a particular Aristotelian necessity: "humans need food". Without further specification, the categorical "consumes food for nutrition" does not allow us to say why Angela behaves irrationally by devoting all her nutritive efforts to swimming in the Atlantic, trying to catch krill.

To save Foot's theory, it would have to be shown that there is an intermediate degree of specificity at which Aristotelian categoricals can be stated. They must be specific enough to normatively guide species-typical actions, yet also vague enough to ensure that violating a categorical will make flourishing impossible. I am deeply sceptical of the potential for such intermediate specification. What statement could possibly constitute an intermediate degree of specification between the categoricals "lives in nomadic bands of hunter-gatherers" and "lives in small agricultural communities", for example? The most obvious answer is, of course, "lives either in nomadic bands of hunter-gatherers or in small agricultural communities". But this is not an option, at least prior to the actual establishment of agriculture. This is because Aristotelian categoricals are generated by how a species *actually* lives, and not by the ways in which it *could* live. That is why it is illegitimate to say of humans living in the Pleistocene that they "dwell either on dry land or in underwater cities". So, it remains the case that the first humans to begin living in agricultural communities thereby violated one of their Aristotelian categoricals.

Foot's account, I contend, is unable to be revised so as to enable it to accommodate the phenomenon of adaptive behavioural change. As a result, there is only a contingent connection between acting in accordance with one's Aristotelian categoricals and flourishing. This shows that

[81] See Wrangham (2009) Chapter 3.

the notion of an Aristotelian categorical cannot be used to support Foot's *eudaemonist* theory of value, and her analysis of the virtues in terms of those categoricals therefore also fails.

6. A Darwinian Alternative to Foot's Approach

The objections discussed in the previous section take issue with a number of features of Foot's account. Specifically, they stem from concerns relating to her a-historical, welfarist theory of function, and her claim that it is possible to identify a species-specific form of human flourishing. Yet what if these objections could be avoided by leaving out the problematic aspects of Foot's account? Would some other, more Darwinian, analysis of virtue ethics then be defensible? In this section I briefly examine such an approach, defended by Jonathan Haidt and Craig Joseph. As will be seen, Haidt and Joseph's account of the virtues shares some of the crucial aspects of Foot's theory. Yet, crucially, Haidt and Joseph's account is offered as an empirical analysis, and is not intended to act as the foundation for a normative theory. Nor does it seem compatible with the sort of robust realism which Foot's theory tries to deliver. As it stands, then, Haidt and Joseph's account ought not to be thought of as the basis for a realist reply to Joyce.

(i) Continuities with Foot's Theory

Haidt and Joseph explicitly endorse a virtue-ethical approach to moral psychology, claiming that "virtue theories are the most psychologically sound approach to morality".[82] According to the account which they defend:

> [t]o possess a virtue is to have disciplined one's faculties so they are fully and properly responsive to one's local sociomoral context. To be kind, for example, is to have a perceptual sensitivity to certain features of situations, including those having to do with the well-being of others, and to be sensitive such that those features have an appropriate impact on one's motivations and other responses. [...] A virtuous person is one who has the proper automatic reactions to ethically relevant events and states of affairs [...].[83]

This characterisation bears a close resemblance to Foot's description of the virtuous agent as one for whom "certain considerations count as reasons for action".[84] In addition to this characterisation, Haidt and Joseph's discussion of the virtuous agent shares certain other important details, both with Foot's account and with the tradition of virtue ethics more generally. For example, Haidt and Joseph take their position to be continuous with "[o]ne of the central tenets of virtue theory",[85] according

[82] Haidt and Joseph (2004) p. 62.
[83] Haidt and Joseph (2004) p. 61.
[84] Foot (2001) p. 12. Emphasis removed.
[85] Haidt and Joseph (2004) p. 62.

to which "the virtues are acquired inductively, that is, through the acquisition [...] of many examples of a virtue in practice".[86] Furthermore, Haidt and Joseph take their approach to be broadly continuous with the claim that the virtues can be:

> defined [...] by reference to universal features of human beings and their environments that combine to define spheres of human experience in which we make normative appraisals of our own and others' conduct [...].[87]

This claim is conceptually close to Foot's notion of the normative evaluation of an organism in terms of its natural goodness.

Clearly, then, Haidt and Joseph's account has much in common with Foot's theory. Yet Haidt and Joseph do not share Foot's welfarist conception of function, and their discussion of the virtues is not predicated on the virtues making some contribution to an agent's *eudaimonia*. As seen in the previous section, these are deeply problematic aspects of Foot's account. Does Haidt and Joseph's discussion therefore suggest a way to supplement Foot's theory, thus enabling it to meet the objections raised in the previous section? In fact it does not. Before we can see why this is the case, however, it will first be necessary to provide a more detailed exposition of Haidt and Joseph's analysis of the virtues.

(ii) *The Nature of the Virtues* Redux

According to Haidt and Joseph, the virtues develop via the influence of evolved moral intuitions. These intuitions constitute an "innate preparedness to feel flashes of approval or disapproval toward certain patterns of events involving other human beings".[88] That is, moral intuitions comprise "little more than flashes of affect when certain patterns are encountered in the social world".[89] Haidt and Joseph describe these intuitions from the perspective of evolutionary psychology. Accordingly, they take them to be the product of specific mental modules, which evolved in response to selection pressures created by the adaptive challenges and problems associated with human beings' social existence. The modules which generate our moral intuitions are thus a "part of the factory-installed equipment that evolution built into us to solve those recurrent problems".[90] These moral intuitions are generated by specific features of various situations, and their modularity means that they can be so generated without the need for any conscious recognition of the situational features *by which* they are generated. The operation of these intuitions is by no means restricted to specifically moral assessments, however. Rather, Haidt and Joseph claim that:

---

[86] Haidt and Joseph (2004) p. 62.
[87] Haidt and Joseph (2004) p. 63.
[88] Haidt and Joseph (2004) p. 56.
[89] Haidt and Joseph (2004) p. 63.
[90] Haidt and Joseph (2004) p. 56.

> most social cognition occurs rapidly, automatically, and effortlessly – in a word, intuitively – as our minds appraise the people we encounter on such features as attractiveness, threat, gender, and status.[91]

Moral intuitions have an important role to play in the generation of our moral judgements according to Haidt and Joseph, who therefore downplay the role of conscious deliberation in moral judgement. They argue that the judgements which people form when presented with hypothetical moral dilemmas "mostly emerge from the intuitive system: people have quick gut feelings that come into consciousness as soon as a situation is presented to them".[92] Haidt and Joseph cite evidence in support of this claim by appealing to the phenomenon of "moral dumbfounding". This phenomenon occurs when subjects, asked to justify their moral judgement of a hypothetical scenario, are unable rationally to do so. This inability does not, however, result in a change in the subject's moral judgement. Rather, subjects simply insist that *something* about the situation as described to them makes it (say) morally wrong. Thus, Haidt and Joseph argue that when focussing "on the reasons people give for their judgments, you are studying the rational tail that got wagged by the emotional dog".[93]

Haidt and Joseph go on to argue that moral intuitions are typically elicited by social situations relating to issues of suffering and compassion, reciprocity and fairness, and the maintenance or transgression of social hierarchies. Each of these issues, they argue, will have presented "long-standing adaptive challenges"[94] to our ancestors. Given the evolutionary importance of meeting these challenges, the moral intuitions which evolved to do so will be species-typical. That is, they will be found universally, across the entire range of human societies.

It is important here to note Haidt and Joseph's claim that the possession of moral intuitions is not to be identified with possession of the virtues. This would have the implausible implication of making every human agent *de facto* virtuous. However, whilst "a flash of intuition is not a virtue [...] it is an essential tool in the construction of a virtue".[95] What, then, *is* a virtue according to Haidt and Joseph's account? The answer to this question is that "[v]irtues are social skills".[96] More specifically, though perhaps rather less informatively, virtues are "characteristics of a person that are morally praiseworthy".[97] Furthermore, these virtues "are social constructions".[98]

We can unpack these remarks to say that, according to Haidt and Joseph's account, the virtues are socially learned evaluations of psychological dispositions and action-types. These evaluations are based on innate, species-typical moral intuitions, which in turn are generated by evolved mental modules. These intuitions, whilst universally shared, are capable of being culturally elaborated upon so as to generate various divergent and "incommensurable moralities".[99] This process of cultural elaboration can be achieved by various means. For example, through story-telling

---

[91] Haidt and Joseph (2004) p. 57.
[92] Haidt and Joseph (2004) p. 57.
[93] Haidt and Joseph (2004) p. 57.
[94] Haidt and Joseph (2004) pp. 58 – 59.
[95] Haidt and Joseph (2004) p. 63.
[96] Haidt and Joseph (2004) p. 61.
[97] Haidt and Joseph (2004) p. 61.
[98] Haidt and Joseph (2004) p. 56.
[99] Haidt and Joseph (2004) p. 56.

children are encouraged to develop sympathetic responses towards the protagonist, often as a result of her possession of character traits explicitly identified, either by the story-teller or by narrative devices, as morally praiseworthy.[100] In addition, reliance on certain specific sorts of moral intuitions rather than others may be socially encouraged. For example, Haidt and Joseph claim that:

> American Muslims and American political conservatives value virtues of kindness, respect for authority, [and] fairness [...]. American liberals, however, rely more heavily on virtues rooted in the suffering module (liberals have a much keener ability to detect victimization) and the reciprocity module (virtues of equality, rights, and fairness).[101]

Lastly, particular virtues may be grounded in different moral intuitions, depending on the culture in which they are found. Loyalty, for example, can be directed primarily at an agent's friends and family, in which case moral judgements about loyalty are likely to be elicited by intuitions which are sensitive to considerations of reciprocity and fairness. However, loyalty may also be valued in terms of a disposition to obey one's superiors, and not to ignore the wishes of individuals higher up the social ladder than oneself. Moral judgements about this form of loyalty will be sensitive not to intuitions pertaining to reciprocity, but rather to those which are elicited by the norms of social hierarchy.[102] Thus particular virtues may have "different eliciting conditions and different appropriate behaviors and responses".[103]

As can be seen from this discussion, there are substantial differences between Haidt and Joseph's evolutionary analysis of the virtues, and the account provided by Foot. In the next sub-section I briefly highlight those differences which are most relevant to the purposes of this chapter. As will be seen, the positions occupied by Foot and by Haidt and Joseph are fundamentally irreconcilable. Furthermore, the latter's Darwinian approach to virtue ethics seems *prima facie* unable to provide the basis for a realist reply to Joyce.

(iii) Discussion

Whereas for Foot the virtues are actions in accordance with practical rationality, For Haidt and Joseph they are social constructions rooted in non-rational affects. Foot is therefore able to put some normative pressure on the immoral agent by accusing her of acting in a practically irrational way, yet there is no scope for doing this in Haidt and Joseph's account. Theirs is an entirely non-normative model of the virtues, as is acknowledged elsewhere by Haidt. The Haidt and Joseph model is derived from Haidt's "social intuitionist" account of moral judgement, which he expressly identifies as making:

---

[100] See Haidt and Joseph (2004) p. 63.
[101] Haidt and Joseph (2004) p. 64.
[102] See Haidt and Joseph (2004) p. 64.
[103] Haidt and Joseph (2004) p. 64.

a descriptive claim, about how moral judgments are actually made. It is not a normative or prescriptive claim, about how moral judgments *ought* to be made.[104]

However, it has been objected that Haidt's social intuitionism does not pay sufficient attention to the role played by reason in the development of our moral judgements. For example, Saltzstein and Kasachkoff argue that on Haidt's model, changes in our moral judgements are produced:

> as a result of either psychological coercion or non-rational compliance. The one avenue of change he does *not* seem to recognise is that brought about by reasoned argument. But there is ample evidence that reasons can be motivating. We are often moved to change our minds and sometimes alter our behavior in response to the reasoned exhortations and arguments of others.[105]

If a criticism of this sort can successfully be made in response to Haidt's view, it might thereby allow his position to accommodate the claim that virtuous actions are in some sense rational, and therefore normative. In fact, however, Saltzstein and Kasachkoff's objection has been convincingly rebuffed by Haidt. He argues that their objection is based on a misreading of his position, which he then goes on to clarify as follows:

> Ordinary people do not spontaneously look for evidence on both sides of a judgment question. But moral reasoning *does* play an important causal role once it is seen as a social activity rather than as a solitary activity. People engage in moral reasoning not so much to figure things out for themselves, in private, but to influence others. [...] Other people's reasons in turn can influence us, and not just by conformity pressure. [...] The social intuitionist model is called "social" precisely because of the importance that social processes play in moral judgment.[106]

Contrary to the objection made by Saltzstein and Kasachkoff, then, Haidt *does* give reasoned debate a part to play in the modification of our moral intuitions. Yet its role is not sufficient for the generation of normativity. Reasoned debate can generate non-coerced moral *agreement*, i.e., it can ensure that we all respond to the same social situations in the same way; but this is a far cry from anything like the claim that reasoned debate can identify moral *truth*. Perhaps, for example, a sufficiently skilled rhetorician would be able to manipulate a society's intuitions in ways that prompted all of its members to agree with vehemently racist moral judgements. There is nothing in Haidt and Joseph's theory of the virtues which would allow such judgements to be identified as false, or which would provide grounds for preferring non-racist judgements to racist ones.

Haidt and Joseph's account of the virtues is preferable to Foot's because it is more plausible, avoiding as it does the deeply problematic issues associated with Foot's theory of function and her commitment to *eudaimonism*. Yet it was precisely these aspects of Foot's theory which allowed her to give a normative, as opposed to a merely descriptive, account of the virtues. The notion of human flourishing also allowed Foot to ground the normative aspect of her theory in a framework that could claim enough objectivity to answer Joyce's sceptical challenge, thereby undermining his

---

[104] Haidt (2001) p. 815. Emphasis added.
[105] Saltzstein and Kasachkoff (2004) p. 275.
[106] Haidt (2004) pp. 284 – 285.

argument for an error theory. Again, this is an attraction which Haidt and Joseph's model lacks. Their account of the virtues sees them as developing from evolved dispositions of evaluation, and not, as in Foot's theory, as propositions capable of truly identifying that which constitutes a "good" organism. Neither of the approaches to virtue ethics discussed in this chapter, then, has the potential to provide a realist reply to Joyce.

7. Conclusion

In this chapter I have critically discussed Foot's argument for normatively evaluating human actions in terms of their contribution to natural goodness. If it were philosophically defensible, Foot's theory would constitute a realist reply to Joyce's argument for moral error theory.

According to Foot, when we judge an action to be morally good we are making the same sort of judgement that we make when we describe the roots of a plant or the colouration of a bird's feathers as good. Both sorts of judgement are predications of natural goodness.

Traits and actions are naturally good insofar as they contribute to an organism's *eudaimonia*. What it means for a particular organism to be *eudaimon* can be empirically determined by observing that organism's species as it exists in its natural environment. Such observation allows us to determine the species' Aristotelian necessities: i.e. those things which the members of the species in question need in order to engage in self-maintaining and reproductive activities. The characteristic ways in which the members of this species go about securing their Aristotelian necessities determine, in turn, the Aristotelian categoricals which can be predicated of the members of that species. Thus, for example, "honey bees perform a waggledance to signify the location of food to conspecifics". Organisms act well to the extent that they act in accordance with their Aristotelian categoricals. This method of evaluation applies to human actions no less than to non-human ones, despite the added complexity of human flourishing. The virtues comprise a core set of human Aristotelian categoricals. Virtuous agency is therefore necessary (though not sufficient) for human flourishing, meaning that action in accordance with the virtues is necessarily practically rational, and that immoral action is necessarily practically *ir*rational.

Foot's theory is deeply problematic. Her welfarist conception of function and commitment to *eudaimonism* generate serious philosophical difficulties, and her attempt to forge a necessary connection between Aristotelian categoricals and an organism's flourishing fails. Foot's claim that moral evaluation is evaluation in terms of natural goodness must therefore be rejected.

An alternative way of characterising the virtues can be found in the work of Haidt and Joseph. Their account has philosophical advantages over Foot's, in that it is not committed to a non-Darwinian conception of function, and nowhere appeals to the notion of *eudaimonia*. Although this alternative is therefore more philosophically plausible than Foot's, however, its plausibility comes at a high price: Haidt and Joseph's account is avowedly non-normative, and so cannot be used to undermine Joyce's sceptical argument for an error theory.

There is, however, still some hope for the moral realist. In the next chapter I discuss Jesse Prinz's argument for a form of realist sentimentalism. As will be seen, Prinz argues that morality is emotionally constituted, and draws support for his argument from Haidt's social intuitionist model of moral psychology. Prinz claims that although our moral judgements are emotionally constituted, and although they must in an important sense be regarded as relative, it is still necessary to posit the existence of moral properties. Prinz therefore characterises his position as a form of moral realism. If this claim proves philosophically defensible, then Prinz's theory could be the best perspective from which to take a realist stand against Joyce.

Chapter Five

Realist Sentimentalism

1. Introduction

In this chapter I critically assess Jesse J. Prinz's argument for a realist form of moral sentimentalism. I begin, in §2, by highlighting the extent to which Prinz's theory is influenced by the philosophy of David Hume, and by explaining how this influence should affect the way in which we read Prinz. Specifically, for example, it allows us to see how Prinz's commitment to empiricism limits the scope of his metaethical project.

In §3 I provide some background detail regarding Prinz's theory of emotion. As will be seen, Prinz takes moral judgements to be emotionally constituted. Given this, an understanding of how he conceives emotion greatly improves comprehension of his metaethical theory. Such an understanding will also generate additional scope for critical reflection on Prinz's theory: this will prove particularly to be the case as regards his claim that emotions are non-cognitive embodied appraisals.

§4 gives a detailed discussion of Prinz's realist sentimentalism. There, I explain Prinz's reasons for holding that moral judgements are emotionally constituted, and show why Prinz takes moral sentiments to be dispositions of emotional reaction. I give an account of Prinz's conception of moral realism, and explain how he is able to argue that moral judgements are emotionally constituted, relativistic, and yet also truth-apt. I also explain the role played by reason in Prinz's metaethics, and show that his account of moral realism satisfies the general desiderata for a realist reply to Joyce.

Finally, in §5 I consider objections to Prinz's account. There I argue that Prinz's philosophy of emotion is inconsistent with his metaethical thesis, and that it is not possible to make these theories consistent without some major conceptual revision. I also argue that Prinz's realist aspirations fail to live up to the standards set for realist theories of morality by other approaches to sentimentalism. Ultimately, then, Prinz's theory cannot be used to form the basis of a reply to Joyce. However, there is still much of use that can be salvaged from Prinz's theory. This will be seen in the next and final chapter, where I draw on some of the aspects of Prinz's account to defend a care-ethical form of realist sentimentalism.

2. Prinz's Humeanism

This section highlights the extent to which Prinz's theory is influenced by earlier work in the sentimentalist tradition, particularly that of Hume. Drawing attention to this influence is helpful

when approaching Prinz's work, as it allows the reader to identify more clearly the conceptual and methodological concerns which are at work in the background of Prinz's philosophy.

In §2.i I give a very general outline of how Prinz takes his work to be a modernisation of and expansion upon Hume's *Treatise*. I identify two important aspects of Hume's empiricism, both of which are to be found in the work of Prinz. These are Hume's commitment to an empirical methodology for philosophical research, and his rejection of psychological nativism. §2.ii claims that Prinz's anti-nativism can be seen as a product of his commitment to Humean empiricism, and that it is this commitment which motivates his scepticism of evolutionary accounts of human behaviour. In §2.iii I show how Prinz's empiricism influences his metaethical theory, making its central concern the explanation of moral judgement, and how this puts Prinz's account in tension with other sentimentalist theories. Lastly, in §2.iv, I explain how Prinz's account of moral realism could, if successful, be used to form the basis of a reply to Richard Joyce's argument for an error theory.

(i) *A Twenty-First Century* Treatise

In *The Emotional Construction of Morals*, Prinz defends a realist form of moral sentimentalism. The sentimentalist tradition has a long history, beginning in the eighteenth century with the work of philosophers such as Hume and Frances Hutcheson. Prinz characterises the theory which he defends expressly in terms of sentimentalism's history, claiming that his account is one which "builds on the ideas developed by Hume and some of his contemporaries".[1] In saying this, however, Prinz does not mean only to acknowledge an intellectual debt to Hume. Rather, Prinz conceives of his overarching philosophical project as one which closely parallels the three books of Hume's *A Treatise of Human Nature*. Thus, Prinz's discussion of moral sentimentalism is intended to supplement his earlier books on the nature of cognition[2] and on the psychology of emotion.[3] Taken together, these three volumes cover precisely the same theoretical ground as the *Treatise*, even treating each subject in the same order as in Hume's earlier discussion. Accordingly, Prinz describes these volumes as "a tribute and modest extension of Hume's masterwork".[4]

It is helpful to draw attention to the close connection between Prinz's work and that of Hume, and not only because it allows us to situate the former in the historical context of a particular philosophical debate. Firstly, doing so prompts a non-isolated reading of Prinz's moral sentimentalism: to fully understand his moral theory, it is necessary also to understand the theory of emotion which Prinz defends. Secondly, the close connection between Hume's moral theory and Prinz's suggests that paying some attention to the former may help us to further illuminate the latter. Such is the task of this section.

In his introduction to the *Treatise*, Hume claims that:

---

[1] Prinz (2007) p. vii.
[2] Prinz (2002).
[3] Prinz (2004).
[4] Prinz (2007) p. vii.

> 'tho we must endeavour to render all our principles as universal as possible, by tracing up our experiments to the utmost, and explaining all effects from the simplest and fewest causes, 'tis still certain we cannot go beyond experience; and any hypothesis, that pretends to discover the ultimate original qualities of human nature, ought at first to be rejected as presumptuous and chimerical.[5]

Hume's comments here are indicative of his commitment to an empiricist methodology. This empiricism manifests itself in Hume's thought in two ways. Firstly, and as seen in the previous quotation, Hume limits the explanatory resources available to any philosophical theory to those which can be directly observed and tested. To attempt to go beyond this limit is simply to introduce an unacceptable degree of speculation into one's philosophical analyses.

The second aspect of Hume's empiricism is to be found in his insistence that the contents of the mind must be understood as being dependent upon an agent's experiences. Thoughts and concepts are analysable either as "simple ideas", which are generated by and correspond exactly to the contents of sensory experience, or as "complex ideas", which are conjunctions of distinct, previously acquired simple ideas. This feature of Hume's thought is often referred to as the "Copy Principle", and it follows from it that appropriate sensory experience is a prerequisite condition for the possession of any thought or concept. Thus Hume writes that "all our simple ideas in their first appearance are deriv'd from simple impressions, which are correspondent to them, and which they exactly represent".[6] To hold this view is to subscribe to what Steven Pinker has termed the "blank slate" theory of mind. According to such a view, any differences between the contents of two agents' minds must be explained by their different life histories. The various experiences comprising these histories are responsible not only for shaping these agents' beliefs, but also the ideas which it is possible for each agent to formulate. As Pinker puts it:

> differences of opinion arise not because one mind is equipped to grasp the truth and another is defective, but because the two minds have had different histories.[7]

Both of these aspects of Hume's empiricism, i.e. his claims about the methodological constraints on inquiry and about the nature of thought, can help us to understand different features of Prinz's philosophy.

(ii) Blank Slates

As seen in Chapter Two, Prinz is deeply sceptical of evolutionary interpretations of human behaviour. This scepticism is the product of Prinz's failure to distinguish between the claim that a trait has evolved, and the claim that a trait is innate; i.e. that it is developmentally canalised. I have already argued that this is a mistake on Prinz's part, and will not repeat that argument here. By connecting Prinz's anti-nativism with his Humeanism, however, it becomes possible to construe it as

---

[5] Hume (1978) p. xvii.
[6] Hume (1978) I i 1. Emphasis removed.
[7] Pinker (2002) p. 5.

stemming from a robust, Humean form of empiricism. Thus Prinz approaches the human mind starting from the assumption that its contents are entirely learned. He concedes that this assumption will not always prove to be true, but argues that one is only licensed to give it up on a case-by-case basis, and in the face of overwhelming evidence to the contrary. Thus Prinz claims that "[t]he best strategy for showing that something is innate is to show that it couldn't be learned".[8]

It is worth briefly noting here that Prinz's rejection of evolutionary analyses of human behaviour, being based on Humean considerations, raises questions about the extent to which Hume would have been in favour of evolutionary approaches to ethics. It is sometimes claimed that certain features of Hume's ethical thought make him a potential ally for the evolutionary ethicist. For example, Michael Ruse writes that:

> [Hume's] is a naturalistic approach to ethics that the Darwinian understands and appreciates. Apart from anything else, his willingness to merge the animal and the human [...] strikes a note that is sweet music to the ears of today's biologist.[9]

However, Prinz's rejection of the evolutionary perspective on Humean grounds should prompt us to question how happy an alliance this really would be. This is not a question which I shall pursue further here, however.

(iii) The Theoretical Primacy of Moral Psychology

As a result of Hume's methodological empiricism, the emphasis of his moral sentimentalism is placed very much on what we would now term moral psychology. As James Baillie observes:

> Hume consciously attempts to do for moral subjects what Newton did for the natural world, namely to provide an accurate classification of mental phenomena, and the principles underlying their activity.[10]

The same assessment is to be found in J. L. Mackie's assertion that Hume's motivation:

> is a demand for an explanation of the sort typically given by the empirical sciences [...] it is an attempt to study and explain moral phenomena (as well as human knowledge and emotions) in the same sort of way in which Newton and his followers studied and explained the physical world.[11]

The significance of the empirical limitations of Hume's inquiry suggests that he was not principally concerned with providing a normative ethics; rather, his main interest lay in describing how and why people make moral judgements as they do, and not in prescribing how they *ought* to make them. This interpretation can be anachronistically glossed as the view that, for Hume, moral psychology is of primary theoretical importance to the moral philosopher.

---

[8] Prinz (2004) p. 128.
[9] Ruse (1990) p. 70.
[10] Baillie (2000) pp. 11 – 12.
[11] Mackie (1980) p. 6.

Reading Hume in this way has the additional attraction of cohering with the claim, generally accepted as having been endorsed by Hume, that it is impossible to validly deduce any normative conclusion from empirical data alone. This claim has come to be known as "Hume's Law". In his statement of this claim, Hume notes that:

> [i]n every system of morality, which I have hitherto met with, I have always remark'd, that the author proceeds for some time in the ordinary way of reasoning, and establishes the being of a God, or makes observations concerning human affairs; when of a sudden I am surpriz'd to find, that instead of the usual copulations of propositions, *is* and *is not*, I meet with no proposition that is not connected with an *ought* or an *ought not*. This change is imperceptible, but is, however, of the last consequence. For as this *ought* or *ought not* expresses some new relation or affirmation, 'tis necessary that it shou'd be observ'd and explain'd; and at the same time that a reason should be given, for what seems altogether inconceivable, how this new relation can be a deduction from others, which are entirely different from it.[12]

Precisely how law-like a claim Hume took himself to be making here has been the subject of some debate. Mackie notes that, coming as it does at the very end of the section in which it appears, the quoted passage was "plainly an afterthought for Hume himself [...] [albeit] an important one".[13] It has been suggested that Hume merely wished to call attention to the need to explicate the move from is-type statements to ought-type statements, without denying that such a move can ever legitimately be made. This interpretation is attractive in the light of arguments designed to refute Hume's Law. Such arguments hold that there are various contexts in which ought-type statements *do* in fact follow from is-type statements, as in the case of hypothetical imperatives and certain types of speech act.[14] However, it is outside of the scope of this chapter to attempt to adjudicate on this issue. Rather, the task at hand is to outline where Prinz stands on the issues of the theoretical primacy of moral psychology and the validity of Hume's Law, and to explain how this affects his sentimentalist account of morality.

In characterising his work as an extension of Hume's, Prinz ought to be read as accepting the theoretical primacy of moral psychology, and as thereby setting a restriction on his analysis of morality. This restriction means that the task Prinz sets for himself is primarily one of describing the moral judgements which we typically make, and of explaining why we do so. As will be seen, Prinz resists the temptation to claim that it is possible for the moral philosopher to identify a set of independently-existing, inter-personal moral norms.

Yet Prinz's focus on the empirical examination and explanation of moral phenomena creates a tension between his account of Humean sentimentalism and other exponents of that view. For example, Justin D'Arms and Daniel Jacobson argue that:

> Humean sentimentalists need not, and we think should not, be seen as ratifying the edicts of whatever moral feelings one happens to have. Sentimentalism is not an epistemological doctrine; instead the fundamental claim is metaethical. [...] Hence,

---

[12] Hume (1978) III i 1.
[13] Mackie (1980) p. 61.
[14] See Mackie (1980) p. 62 for discussion.

> philosophers who take inspiration from Hume must allow reasoning, as well as feeling, to play a role in evaluative judgment. The central challenge for sentimentalism is to preserve the idea that values are somehow grounded in the sentiments, while at the same time making sense of the rational aspects of evaluation.[15]

These differing conceptions of what sentimentalism ought to do seem, at least *prima facie*, to constitute an irreconcilable difference between Prinz's version of sentimentalism and that of D'Arms and Jacobson. At the end of this chapter I will consider whether these differences are fundamental or merely apparent.

A useful way to explore this issue is to confront Prinz's attitude towards Hume's Law. This attitude is one of qualified acceptance. Prinz holds that "Hume's Law is basically true".[16] However, Prinz also asserts that "[t]he metaethical theory and moral psychology that I will be defending in the chapters that follow offers a way to cross the is/ought boundary".[17] That is, Prinz holds that a sufficiently precise descriptive account of how we make moral judgements will have genuinely normative implications for moral practice. As Prinz puts it, "we can figure out what our obligations are by figuring out what our moral beliefs commit us to".[18] This claim might seem to move Prinz closer to the position endorsed by D'Arms and Jacobson. However, the empirical concerns which underlie Prinz's theory preclude the possibility of an inter-subjectively prescriptive normative ethics. Thus, *contra* D'Arms and Jacobson, Prinz holds that to adopt a Humean sentimentalism inevitably "leads to moral relativism".[19]

(iv) Moral Realism in Prinz and Hume

Despite Prinz's commitment to relativism, he describes his view as a vindication of moral realism. In this he might be thought substantially to depart from Hume's view, which is sometimes construed along anti-realist lines as a precursor of A. J. Ayer's emotivism. Interpretations of this sort are problematically anachronistic, however, and risk illicitly projecting much later metaethical debates into Hume's thought. Indeed, Hume can plausibly be read as a moral realist. As James Baillie argues:

> Hume regards ethical discourse itself as perfectly legitimate. That is, to *employ* moral concepts like right and wrong is not, in itself, to be guilty of metaphysical error.[20]

In endorsing moral realism, then, Prinz need not be seen as departing from Humean orthodoxy.

Prinz's brand of moral realism makes his theory particularly relevant to the purposes of this thesis. As will be seen, Prinz endorses a form of causal realism. On this view, moral properties are

---

[15] D'Arms and Jacobson (2000) p. 722.
[16] Prinz (2007) p. 7.
[17] Prinz (2007) p. 1.
[18] Prinz (2007) p. 1.
[19] Prinz (2007) p. viii.
[20] Baillie (2000) p. 15. Original emphasis.

properties of actions, or of situations, which cause us to experience a moral emotion. Prinz's realism does not specify that these properties need themselves be conceived of as intrinsically motivational. By identifying the formation of a moral emotion with the making of a moral judgment, Prinz is able to explain moral motivation by directly appealing to facts about an agent's psychology, rather than to the nature of moral properties. Prinz's account does not therefore pose any serious problems for naturalistic realism. The properties of moral rightness and wrongness are, according to his theory, no more difficult to naturalise than are the properties of fearfulness or disgustingness. Joyce's objection that moral properties are impossible to satisfactorily naturalise would therefore have been disarmed.

By conjoining this form of realism with the other aspects of his theory, Prinz's account suggests a way in which to incorporate a normative element into the moral psychology outlined by Jonathan Haidt and Craig Joseph, discussed in the previous chapter. If Prinz is right, then it is possible to accept Haidt and Joseph's relativistic social intuitionism, whilst also claiming that moral properties exist. This is an enticing prospect, as it was the absence of a normative element in Haidt and Joseph's account which unfitted it to act as a realist reply to Joyce's argument for an error theory. If Prinz's account of moral properties is correct, then, it would show that even if Haidt and Joseph are right about the intuitive basis of moral judgement, it is still possible to talk legitimately about the truth and falsity of certain moral beliefs (at least up to a point). Not only does Prinz's theory point to a way of meeting Joyce's scepticism about the merits of naturalism, then; it is also compatible with a Darwinian moral-psychology of the virtues. If successful, Prinz's theory could happily sit alongside Haidt and Joseph's. Taken together, these theories would then constitute a normative virtue ethic.

It is worth briefly noting here that Prinz does not take his account to be an attempt to vindicate any form of virtue theory. He rejects virtue ethics for two main reasons, each of which constitutes a denial of the claim that the virtues are dispositions which it is "natural" for humans to possess. The first of these reasons is that conceptions of well-being (i.e. of *eudaimonia*) are culturally constructed, and so no account of flourishing is more plausibly "natural" or universal than another. Prinz's second reason for rejecting the naturalness of the virtues is that virtuous traits will often not be genetically adaptive, and are therefore unlikely to have been hard-wired by natural selection.[21] I will not give a detailed discussion of these issues here. Rather I will simply point out that, whilst Prinz's concerns apply to virtue ethical theories akin to that of Foot, they do not apply to Haidt and Joseph's alternate account of the virtues. It is therefore doubtful that Prinz would have any principled objection to his theory being used to add a normative dimension to the latter.

I will ultimately conclude that Prinz's account of moral realism fails. Nevertheless, there are numerous aspects of Prinz's theory which can be incorporated into a more successful, care-ethical form of realist sentimentalism, which I go on to defend in the next chapter. Specifically, Prinz's approach to Hume's law; his discussion of basic values; and his insight that Haidt and Joseph's moral psychology is compatible with moral realism will all have a place in the account which I defend. Yet this chapter is not solely devoted to setting up the next. Paying attention to Prinz's conception of moral realism helps to identify some of the desiderata which realism must satisfy. It will suggest the view that the best way of judging whether a moral theory can appropriately be described as realist

[21] See Prinz (2007) pp. 157 – 158 for discussion.

will depend on what it is possible normatively to *do* with that theory; i.e. in the theory's potential to be used to resolve moral debate.

Before critically assessing Prinz's attempt to vindicate moral realism, however, it is first necessary to discuss his theory of emotion. As previously noted, this theory forms the basis of Prinz's analysis of moral psychology. Any attempt fully to understanding the latter must therefore begin with an account of the former. Such an account is given in the next section.

3. Prinz's Theory of Emotion

For Prinz, "emotions constitute moral judgments".[22] But if moral judgements are emotions, it is reasonable to ask, what then are emotions? The answer to this question, at least according to Prinz, is that emotions are the perception of bodily changes. That is, "an emotion is an inner state that registers a pattern of change in the body".[23] More specifically, emotions are non-cognitive, embodied appraisals.

In this section, I explain precisely what Prinz means by this. I begin, in §3.i, by distinguishing cognitive from non-cognitive theories of emotion. §3.ii gives an account of Prinz's notion of an embodied appraisal, according to which emotions represent "organism-environment relations with respect to well-being".[24] This account is then further unpacked in §3.iii by an elaboration of the concept of representation as applied to emotion. Finally, §3.iv discusses Prinz's distinction between basic and higher-cognitive emotions. This discussion is of central importance to an understanding of Prinz's conception of the moral emotions, as he argues that these are typically constituted by higher-cognitive emotions.

(i) Emotions are Non-Cognitive

By insisting that emotions are non-cognitive, Prinz rejects the view that the cognitive assessment (or appraisal) of a situation plays a necessary role in the formation of an emotion. It is important to note that in saying this, Prinz does *not* mean to deny that cognition has the capacity to *influence* our emotional responses. Thus he writes that:

> [a] thousand everyday experiences teach us that thoughts can effect emotions. If you see a man approaching you from the far end of a street and you form the belief that he is an old friend, you will feel happy, but if you form the belief that the man is a stranger

[22] Prinz (2008a) p. 162.
[23] Prinz (2005) p. 15.
[24] Prinz (2004) p. 52.

> and a possible assailant, you will feel fear. Positive thoughts lead to positive affect, and negative thoughts lead to negative affect.[25]

Rather than claiming that cognitive judgement is unable to *influence* emotion, Prinz argues that cognition is not an essential *feature* of emotion. He supports this claim by endorsing the argument, made by Robert Zajonc, that "emotions are phylogenetically and ontogentically prior to cognitions".[26] To say this is simply to claim (i) that emotions occur in species which do not possess the cognitive capacities found in human beings; and (ii) that the emotions develop in human infants prior to their capacity for anything like adult human cognition.

Prinz also draws on empirical data from facial feedback studies in support of his rejection of cognitive theories of emotion. These studies purport to show that a subject's facial expression has the ability to influence how that subject responds to an emotional stimulus. Prinz cites an experiment in which subjects were invited:

> to fill out a questionnaire, [whilst] holding a pen in their mouths. Some subjects were asked to hold the pen with their puckered lips, forming an inadvertent grimace, while others were instructed to hold the pen between their teeth, forcing a subtle grin. Among items in the questionnaire, subjects were asked to rate cartoons. Subjects in the teeth condition rated the cartoons as more amusing than subjects in the lips condition, suggesting that their emotional response was being elevated by their unintended facial expressions.[27]

It is not my intention to offer a detailed philosophical criticism of Prinz's account of emotion; rather, I am primarily concerned with how that account dovetails with his metaethical theory. Nevertheless, it is worth briefly noting at this point that the facial feedback study that Prinz discusses does not provide a truly compelling argument in favour of non-cognitive theories of emotion. A proponent of cognitivism could grant that somatic changes play some role in the formation of emotion, and that such studies highlight that role. The cognitivist could go on to insist, however, that the study which Prinz discusses also shows there to be an essential cognitive element at work in our emotional responses. As Prinz reports the data, subjects found the cartoons amusing to various degrees. None of them are reported as finding the cartoons deeply offensive, or inexplicably melancholy. Had such responses been elicited, the conclusion that certain emotions can be produced *solely* as a result of a change in facial expression would be much more compelling. As it is, however, all of the subjects responded in a way which seems contextually appropriate; i.e. with more or less *amusement*.

Prinz endorses several additional arguments against cognitive theories of emotion, but an exhaustive discussion of these is beyond the scope of this chapter.

(ii) Emotions are Embodied Appraisals

[25] Prinz (2004) p. 30.
[26] Prinz (2004) p. 33.
[27] Prinz (2004) p. 36.

So far then, some of the reasons for Prinz's rejection of cognitive theories of emotion have been noted. Prinz often characterises the act of cognitive assessment as the act of making an appraisal, but it will be recalled that for Prinz emotions are also a type of appraisal; specifically, embodied appraisals. What, then, is the difference between a cognitive appraisal and an embodied appraisal?

According to Prinz, philosophers who subscribe to cognitive theories of emotion typically describe the emotions as cognitive *appraisals*. As previously seen, Prinz rejects the claim that emotions are cognitive. However, he agrees with cognitive theorists that emotions constitute appraisals. To see what this means for Prinz's account, it is helpful to express this difference in terms of three theses, each of which Prinz takes to be central to cognitive theories of emotion. The first of these theses is that an emotion is "a mental state consisting of a representation of a proposition and an attitude toward that proposition".[28] According to the second thesis, emotions are "something above and beyond the bodily changes or inner states that register bodily changes".[29] Finally, the third thesis claims that emotions are appraisals, i.e. that they "are representations of an organism-environment relationship that bears on well-being".[30]

Prinz's non-cognitivism prompts him to reject the first two of these theses. This leaves him with thesis three, which he terms the "appraisal hypothesis". Once it is abstracted from the first two theses, Prinz is happy to accept this claim. He therefore describes his view as one according to which emotions are not *cognitive* appraisals, but are instead *embodied* appraisals. Prinz summarises his conception of emotions as embodied appraisals as the view that:

> emotions necessarily comprise representations of organism-environment relations with respect to well-being [...] [and] that such representations can be inextricably bound up with states that are involved in the detection of bodily changes.[31]

It is possible to get a clearer sense of the bodily changes Prinz has in mind by noting his endorsement of what he terms the "subtraction argument".[32] According to this argument, "emotion phenomenology seems to be exhausted by sensations of bodily changes".[33] Defending this claim in an early passage, Prinz asks the reader to:

> imagine feeling an emotion, and then to imagine systematically subtracting away the feelings of corresponding bodily states. Imagine feeling elated without your heart racing. [...] [O]nce bodily feelings are gone, there seems to be nothing left to the emotional experience.[34]

This is not to imply that for Prinz all emotions are necessarily *feelings*, however. Prinz's emphasis is on *bodily changes*, and he does not insist that such changes are always perceived (i.e.

---

[28] Prinz (2004) p. 22.
[29] Prinz (2004) p. 25.
[30] Prinz (2004) p. 25. Note that Prinz sometimes refers to such organism-environment relations as "concerns".
[31] Prinz (2004) p. 52.
[32] Prinz (2004) p. 56.
[33] Prinz (2004) p. 56.
[34] Prinz (2004) pp. 4 – 5.

felt) by their subject. Thus, in a later article, Prinz argues for the existence of unconscious emotional states; i.e. emotions which are not accompanied by the perception of feelings.

> We can have unconscious visual states, unconscious auditory states, unconscious tactile states, and so on. It seems overwhelmingly likely, then, that we can have unconscious perceptions of the patterned bodily changes that constitute our emotions. [...] For example, imagine being woken up by the sound of glass shattering in your living room. You might assume that burglars are breaking in and attend intensely to the sound. At the very same time, your body will enter into a fear pattern, but you might not experience the fear consciously because attention is consumed elsewhere. After waiting to hear if there is any more noise, you hear your cat scurrying about and you realize she must have knocked over a vase. You then notice, and only then, that your heart is racing, and your breathing is strained [...]. You were afraid, but you didn't realize it.[35]

Still, Prinz does hold that all *conscious* emotions are bodily feelings. Of course, this is only a part of Prinz's account. Nothing has yet been said about precisely how emotions *qua* embodied appraisals can be said to "comprise representations of organism-environment relations with respect to well-being".[36] It is to this issue that I now turn.

(iii) Emotions as Representations

Prinz argues that for something to count as a representation, it must do more than simply convey information. This is because it is part of the concept of a representation (at least for Prinz, who here follows Fred Dretske[37]) that it has the potential to be mistaken. To clarify this position, Prinz compares representations with other forms of information-carrying media. He begins by arguing that:

> [w]e can say that the smoke on the horizon *means* there is a fire [i.e. because the smoke carries information about the presence of fire]. But carrying information is not sufficient for representation. Smoke does not represent fires.[38]

In contrast, however,

> a dog concept is a mental state that is reliably caused by (i.e. becomes active as a result of) encounters with dogs. But a mental state that is reliably caused by dogs may also be reliably caused by wolves, foxes, or well-disguised cats. [...] To count as a dog representation (and not a dog-or-wolf-or-fox-or-cat, representation), there must be a

[35] Prinz (2005) p. 17.
[36] Prinz (2004) p. 52.
[37] See, for example, Dretske (1986).
[38] Prinz (2004) p. 53.

way to say that I am making an *error* when my dog concept activates in response to wolves, foxes and cats.[39]

A dog concept may be considered erroneous when applied to other canids (or very sneaky felids), Prinz argues, because the dog concept is functional: it exists in order to convey information *specifically about dogs*. The smoke produced by fire has no such aetiology; it is simply a by-product of combustion.

> A dog concept is a mental state that is reliably caused by dogs *and* was acquired for that purpose. [...] After such a state is formed, it *carries information* about dogs, foxes, and wolves, because all these things can cause it to activate, but it only *represents* dogs, because it was set up as a result of dog encounters.[40]

It has already been noted that, according to Prinz, emotions are representations of organism-environment relations which pertain to well-being. It can now be seen that for Prinz this means that emotions do not simply convey information; rather, emotions have a specific functional aetiology, and it is this aetiology which accounts for their existence.

It might, at this point, be thought that Prinz is poised to claim that emotions represent the bodily changes by which they are constituted. Yet he does not take this to be the case: "[e]motions clearly 'register' changes in the body, but there is still a further question about what such states represent".[41] Prinz supports this statement by drawing on evolutionary considerations. If emotions had been naturally selected to represent bodily changes, he argues, then it would follow that some adaptive advantage was conferred upon an agent which had the ability to represent to itself, for example, an increase in its heart rate. Prinz persuasively argues that this is an implausible hypothesis. As he puts it, "[i]t is not clear why it is advantageous to know when my blood vessels are constricting".[42]

As an alternative to this view, Prinz suggests that emotions represent specific environmental cues, which they have evolved to detect. Prinz, adapting a concept introduced by Richard Lazarus,[43] categorises these cues as "core relational themes". Prinz offers a lengthy list of core relational themes, which it is unnecessary to reproduce in full; a selection of examples from that list, prefixed by their corresponding emotion, will be sufficient to illustrate the idea. Thus, among his core relational themes, Prinz identifies the following:

Anger: A demeaning offense against me and mine.

Fright: Facing an immediate, concrete, and overwhelming physical danger.

Guilt: Having transgressed a moral imperative.

Shame: Having failed to live up to an ego-ideal.

---

[39] Prinz (2004) p. 53.
[40] Prinz (2004) pp. 53 – 54.
[41] Prinz (2004) p. 58.
[42] Prinz (2004) p. 59.
[43] See Lazarus (1991).

Compassion: Being moved by another's suffering and wanting to help.[44]

The core relational themes which Prinz posits are clearly of evolutionary salience, although the relevance of some such themes is restricted to social organisms. Nevertheless, an organism which was oblivious to the presence of overwhelming physical danger, for example, or a social co-operator who experienced no guilt on failing to reciprocate some altruistic action (therefore being more likely to re-offend) would be at a serious adaptive disadvantage. The suggestion that emotions exist to represent these concerns is therefore evolutionarily plausible.

As we have already seen, it is a condition of a thing's being a representation that it also be capable of *mis*representation. Prinz's account of emotions as the representation of core relational themes allows him to accommodate this condition. This is because, "[r]ather than saying that sadness represents things that are sad to us, we can say that sadness represents loss".[45] For Prinz, there is a fact of the matter about what kind of thing *constitutes* a loss, and this fact applies even if we do not *represent* an appropriately constituted event to ourselves *as* a loss. Prinz further elaborates this argument by appealing to other types of representation.

> Being dangerous, like being poisonous, is a relational property, and a relative property. Something can be dangerous only *to* some creature or other, and whether or not something is dangerous depends on the creature in question. But being dangerous does not depend on being *represented* as dangerous. Radiation would be dangerous even if we didn't know that it is. Fear represents the property of being dangerous even though that property is possessed by some things that we do not in fact fear.[46]

Thus for Prinz, "[w]hile beliefs aim at the True, emotions aim at Relations that Matter".[47] Accordingly, when an emotion fails accurately to represent some aspect of organism-environment relations, it can legitimately be said to *mis*represent them. Prinz is therefore able to conclude that emotions are in fact appraisals. However, Prinz's appeal to the notion of core relational themes as the object of emotional representation points to a possible objection to his account of emotions as non-cognitive states. Prinz's attempt to resolve this issue introduces an important new aspect to his theory of emotion, which has direct implications for his metaethical theory. It is to this objection, and to Prinz's response to it, that I therefore now turn.

(iv) Basic Emotions and Higher Cognitive Emotions

Prinz identifies a possible objection to the non-cognitive aspect of his embodied appraisal hypothesis. According to this objection, many of the items on Prinz's list of core relational themes seem necessarily to involve cognitive attitudes. In particular, moral emotions such as guilt and shame seem to be inextricably bound up with beliefs. This would mean that the content of those moral emotions is not limited to bodily feelings. For example, Prinz suggests that "guilt may be

---

[44] Adapted from Prinz (2004) p. 16.
[45] Prinz (2004) p. 63.
[46] Prinz (2004) pp. 63 – 64.
[47] Prinz (2004) p. 80.

sadness brought on by the belief that one has committed a harmful transgression".[48] If moral emotions (for example) do involve cognitions, then they are not the embodied appraisals which Prinz takes them to be.

One way of accommodating this objection, Prinz argues, is to draw a distinction between basic emotions and higher cognitive emotions. Basic emotions, on this view, are those "from which all others are derived".[49] They are innate, i.e. "present in all normally developing members of the species. [...] They are evolved patterns of response [...] not derived from other emotions".[50] Prinz does not attempt to provide an itemised account of each of these basic emotions, choosing instead to defer this task to future empirical research.

By way of contrast with the basic emotions, Prinz argues that "non-basic", or "higher cognitive", emotions can be conceived of as "blends of basic emotions".[51] So, for example, "[f]eeling sadistic may involve a blend of anger and joy".[52] However, the phenomenology of many emotions does not seem to be adequately captured by this sort of account, as Prinz himself observes:

> [c]onsider romantic jealousy, which may involve ideas of infidelity, sex, and entitlement. It is hard to see how any of these ideas could emerge from basic emotions. None of these ideas are contained within the emotions which are usually catalogued as basic. If emotions were all basic or blends of basic emotions, some of the ideas that constitute romantic jealousy would have to emerge *ex nihilo*.[53]

Thus, the claim that higher cognitive emotions are blends of basic emotions needs supplementing. Accordingly, Prinz goes on to suggest that:

> [a]ppraisals can be purely embodied, or they can include cognitive elaborations of embodied appraisals. Basic emotions are embodied appraisals. Higher cognitive emotions, one might suppose, are either blends of two basic emotions [...] or combinations of basic emotions and cognitive elaborations.[54]

Prinz argues that this view allows him to accommodate aspects of our emotional lives which seem incompatible with an unmodified version of the embodied appraisal hypothesis. These include those situations which elicit a particular emotional response in us as a direct result of our cognitive elaboration of that situation, as opposed to any of its actual features. Thus, "one might become afraid while walking down a desolate city street and recognising that one could be attacked without anyone around to help".[55]

It is important to note here that although Prinz talks of "higher cognitive emotions", he denies that the position outlined above commits him to a cognitive theory *of* emotion. Crucially, in Prinz's account, *higher cognitive emotions are still construed non-cognitively*. How is this possible?

---

[48] Prinz (2004) p. 93.
[49] Prinz (2004) p. 86.
[50] Prinz (2004) p. 88.
[51] Prinz (2004) p. 92.
[52] Prinz (2004) p. 92.
[53] Prinz (2004) p. 93.
[54] Prinz (2004) p. 97.
[55] Prinz (2004) p. 98.

In arguing for this claim, Prinz appeals to a distinction between the notion that higher cognitive emotions are partially *constituted* by cognitions, and the view that they are necessarily *accompanied* by them. He rejects the former view, whilst endorsing the latter. For Prinz, then, "[t]he cognitions that elaborate them [i.e. higher cognitive emotions] are *prior conditions*, not *constituent parts*"[56] of the emotion. Thus higher cognitive emotions (many of which constitute the moral emotions) solely comprise non-cognitive embodied appraisals, as is the case with basic emotions.

The short summary of Prinz's theory of emotion provided in this section has, of necessity, left out a considerable amount of detail. It has also avoided critical discussion of Prinz's views. Whilst a full discussion of the issues raised by Prinz's theory of emotion is desirable, it is also outside the scope of this chapter. The purpose of this section has simply been to add some depth to the forthcoming discussion of Prinz's moral theory. It is to that discussion which I now turn.

4. Prinz's Sentimentalism

In this section, I describe Prinz's brand of realist sentimentalism. As previously noted, Prinz holds that when we make a moral judgement about an event, we are reacting emotionally to that event. Here, Prinz means not only that our moral judgements are necessarily accompanied by certain emotions, nor only that those emotions have the ability to influence our moral judgements (though he in fact subscribes to both of those theses). Rather, Prinz makes the stronger claim that our emotions *constitute* our moral judgements. I begin the section by explaining Prinz's reasons for defending this thesis. In §4.ii I outline Prinz's conception of moral sentiments as emotional dispositions, and his identification of the possession of such dispositions with subscription to a moral rule. Following this outline, §4.iii gives an account of Prinz's analysis of the nature of moral properties. It is this analysis that allows Prinz to claim that moral judgements are capable of being true or false, despite their emotional constitution. Nevertheless, it also follows from Prinz's analysis of moral properties that moral judgements are only true relative to a particular speaker. Lastly, in §4.iv, I explain how Prinz attempts to use our capacity to engage in rational debate to remove some of the philosophical tension generated by his endorsement both of realism and of relativism.

(i) Moral Judgements as Emotionally Constituted

Prinz does not claim that he is able to supply a conclusive argument in favour of the emotional constitution of morality, or "emotionism", as he terms that thesis. Rather, he provides empirical details which lend support to emotionism, and which allow him to conclude that emotionism "makes the most sense of the data".[57]

---

[56] Prinz (2004) p. 98.
[57] Prinz (2006) p. 36.

The first piece of evidence which Prinz cites in favour of emotionism draws on the co-occurrence of emotions with moral judgements. He notes that "every neuroimaging study of moral cognition seems to implicate brain areas associated with emotion".[58] Thus, "[w]hen we do things that violate moral values, we incur emotional costs".[59] However, Prinz concedes that such data is also consistent with the claim that, although we emotionally engage with the upholding of moral norms, those norms are not themselves emotionally constituted.

Prinz's second piece of evidence for emotionism is based on the influence which emotions can exert on our moral judgements. Prinz suggests that the fact that stronger moral sanctions are applied to certain sorts of action is a result of the emotional intensity with which we regard actions of that type. For example, if we find directly killing an individual to be morally worse than passively allowing that individual to die, this is because the former action-type arouses stronger emotions in us. Prinz argues that an emotionist explanation of this sort allows us to account for the different intuitions elicited by moral-philosophical thought experiments.

> Pushing someone in front of a trolley seems wrong, but it seems okay to divert a trolley away from five people and toward one. Why is this? [...] One answer is that [...] [w]e have negative feelings about killing, and positive feelings about saving lives, and few feelings about letting die.[60]

An alternate, cognitivist explanation to emotionism is also available, however: it is a fact that pushing someone in front a trolley is morally worse than directing a trolley away from five people and towards one; and that killing is morally worse than passively letting die. If recognising moral reasons elicits strong emotions in us (because moral reasons are reasons which we care deeply about), it is only to be expected that moral intuitions will be stronger when they relate to situations that exemplify morally worse transgressions. Prinz is aware of this alternative view, and presents a second, related argument about how moral intuitions are emotionally influenced.

> When subjects say that it is morally permissible to pull [a] lever to save five people and kill one, they are imagining that the lever is far away from the tracks. Now suppose we tell subjects that the lever is just a few inches away from the person who would be killed if the lever were pulled. Imagine yourself in that situation. A man is tied down on the tracks right next to you. You cannot free him. He is writhing around and howling in terror. [...] Would you sacrifice the person at your feet? Would that be morally acceptable? Here, I think intuitions would change.[61]

This is a more persuasive argument. Of course, it is possible to deny that the change in intuitions elicited by this scenario would track an objective moral truth about what should be done in such a case. However, it is more difficult to deny that such a change would be brought about by an emotional response to the altered features of the situation. All that Prinz is hoping to establish at this stage of his argument is that emotions influence our moral judgements, and the thought experiment which he outlines strongly supports this claim.

---

[58] Prinz (2007) p. 22.
[59] Prinz (2007) p. 22.
[60] Prinz (2007) p. 24.
[61] Prinz (2007) p. 25.

Prinz then goes on to consider a third piece of evidence in favour of emotionism: the phenomenon of moral dumbfounding (previously discussed in Chapter Four, §6.ii). Moral dumbfounding, it will be recalled, occurs when a subject forms a persistent moral judgement, despite being unable rationally to defend that judgement. Jonathan Haidt provides an example of one such scenario, in which a brother and sister (Mark and Julie) engage in consensual incest. Mark and Julie use a condom, find their experience mutually rewarding, and then vow never to repeat it. Asked whether their action was moral, Haidt found that many subjects said that it was not. Furthermore, subjects persisted in this judgement despite being unable to explain *why* Mark and Julie acted immorally.

> They point out the danger of inbreeding, only to remember that Julie and Mark used [...] birth control. They argue that Julie and Mark will be hurt, perhaps emotionally, even though the story makes it clear that no harm befell them. Eventually, many people say something like, "I don't know, I can't explain it, I just know it's wrong".[62]

For both Prinz and Haidt, the existence of moral dumbfounding suggests that morality is emotionally constituted. Prinz argues that the scenarios presented in moral dumbfounding thought experiments foreground what he terms "basic values". These are values which are not susceptible to any rational analysis or correction. Once they have been elicited, they effectively act as conversation-stoppers. That is, basic values "reveal something about the practise of reason-giving in morality".[63]

> Why is drunk driving wrong? The answer is that it endangers innocent lives. Why is it wrong to endanger? Because danger is risk of harm, and harming an innocent person is wrong. Why is it wrong to harm an innocent person? Here the question becomes odd. [...] At some point the why-question looks misplaced, bizarre, or even depraved. [...] [B]asic values are implemented in our psychology in a way that puts them outside certain practices of justification. Basic values *provide* reasons, but they are not *based* on reasons. [...] Moreover, basic values seem to be implemented in an emotional way. When we get down to basic values, passions rule.[64]

Prinz's final piece of evidence in support of emotionism is based on studies of psychopathy. He claims that:

> [p]sychopaths seem to be the closest thing we have to real-world amoralists. They are perfectly intelligent and articulate. They seem to comprehend moral values, but they are utterly indifferent to them.[65]

Prinz goes on to argue, however, that psychopaths do *not* in fact comprehend moral values. He supports this claim by citing research which purports to show that psychopaths are unable to perceive a difference between moral rules and rules which are the product of social convention.

---

[62] Haidt (2001) p. 814.
[63] Prinz (2007) p. 32.
[64] Prinz (2007) p. 32. Emphasis added.
[65] Prinz (2007) p. 42.

According to the study in question, the psychopath's inability to distinguish between moral and conventional norms shows that for them, moral norms are like the norms of etiquette: "a group of more or less arbitrary conventions that place demands on us only because they have been adopted by a social group".[66] Furthermore,

> [i]n psychopathy a deficit in moral motivation co-occurs with a deficit in moral competence. This suggests that the two are linked. In fact, leading explanations of psychopathy maintain that the deficit in moral comprehension is a direct result of the emotional deficit.[67]

If correct, this account would explain why psychopaths often disregard moral norms. If the force of those norms is grounded in affective responses, and if psychopaths lack the ability to experience or respond to those affects, then they will lose their behavioural salience. For psychopaths, identifying moral norms could sometimes be as problematic as identifying the appropriate piece of cutlery in a complexly arranged dinner service. They lack the emotional responses which make those norms stand out from the crowd of arbitrary conventional norms. This is not to say that all psychopaths are as likely to commit murder as they are to mistake the butter knife for the fish knife. However, they may be consistently less conscious of how their actions may cause distress or suffering to those around them, or to perceive such distress as a reason to change their course of action.

It is difficult to assess the empirical evidence which Prinz draws on in his discussion of psychopathy. It must suffice here to say that if this evidence is correct, it does indeed provide strong support for Prinz's case for emotionism.

As noted above, Prinz does not believe that he has provided a conclusive argument for emotionism. He does, however, assert that the arguments that he makes constitute strong *prima facie* support for emotionism, and that as a result emotionism "is worth exploring".[68] It is to Prinz's exploration of emotionism that I therefore now turn.

(ii) Moral Sentiments as Emotional Dispositions

Prinz distinguishes between what he terms "reactive" and "reflexive" moral emotions. This distinction, he claims, "correspond[s], roughly, to what some authors call 'other-blame' and 'self-blame' emotions".[69]

The reactive class of moral emotions comprises moral anger, moral disgust, and moral contempt. These are all conceived as non-basic emotions, and are therefore importantly different from their non-moral counterparts. For example, Prinz holds that we can be angry at someone

---

[66] Prinz (2007) p. 44.
[67] Prinz (2007) p. 44.
[68] Prinz (2007) p. 49.
[69] Prinz (2007) p. 69.

without taking that individual to have behaved immorally. Anger is only moral anger when it pertains to acts which are either unjust or which violate another's rights.

Moral disgust is somewhat different. Rather than being a blend of basic-emotion-plus-cognitive-eliciting-condition, as in the case of anger and rights violations, disgust is seen as being "metaphorically extended"[70] to include notions of spiritual as well as physical contamination. This spiritual contamination is seen as resulting from a violation of the natural order.

The third and final reactive moral emotion is moral contempt. This is a blend of moral anger and moral disgust, and is "elicited by violations against community".[71] This emotion constitutes a blend of the two other reactive moral emotions because community violations not only infringe upon the rights of other members of the community, but also breach the norms which govern the role each member of the community is assigned (i.e. "the natural order of a human collective"[72]).

Reflexive, or self-directed, moral emotions include guilt, shame, and a blend of these two emotions for which Prinz does not have a name. Each of these is, again, a non-basic emotion, and may be thought of as the reflexive counterparts of moral anger, disgust, and contempt, respectively. Thus:

> [g]uilt may represent the concern expressed by: I have violated an autonomy rule against a member of a group with which I feel a connection. [...] A person who feels shame will typically feel dirty, unworthy, or corrupt. [...] [P]eople who transgress against community will feel a blend of guilt and shame.[73]

Prinz argues that guilt and shame focus on act and agent respectively. That is to say, when an agent feels guilt she views her action as wrong, but without also judging that she is a bad person for performing that action. When experiencing shame, however, she will see herself as having been made worse by her action. In the case of a blend of these emotions, the focus of each is experienced simultaneously.

> If someone discovered your crime you would want to conceal yourself, as if in shame. But you would also dwell on the harm you caused in an act-directed way, as if guilty.[74]

Although Prinz's focus is on negative moral emotions, he does acknowledge that a similar account needs to be provided for positive emotions. This is not something which he attempts to do in any great detail. He does suggest, however, that positive moral emotions may be usefully categorised according to "who is acting, and who is being acted upon".[75] Prinz suggests admiration, dignity, gratitude and gratification as likely candidates for these emotions, although he does not suggest that they constitute an exhaustive list.

Despite spending so long discussing the moral emotions, Prinz does not identify subscription to a moral rule with the possession of a corresponding moral emotion (although moral emotions *do*

---

[70] Prinz (2007) p. 71.
[71] Prinz (2007) p. 74.
[72] Prinz (2007) p. 74.
[73] Prinz (2007) p. 77.
[74] Prinz (2007) p. 77.
[75] Prinz (2007) p. 81.

constitute individual moral judgements). Rather, to subscribe to a moral rule is to have a stable disposition to experience a particular range of emotional responses in relation to certain types of action.

> [W]hen we judge that something is wrong, [...] [that] judgment will be an expression of the underlying emotional disposition. A standing judgment that something is wrong consists in the standing disposition (or its categorical basis), and an occurrent judgment will ordinarily contain a specific emotion that manifests the disposition.[76]

Prinz identifies such dispositions as "sentiments". The various emotional dispositions which each of our sentiments constitutes do not all resemble one another. Some sentiments give rise to numerous emotional states. Thus care is a sentiment, and in caring for someone "you will be delighted by her achievements and distressed by her suffering".[77] Phobias, conversely, "manifest themselves via the same or similar emotions on every occasion."[78]

It is important to note that by identifying subscription to a moral rule with the possession of a sentiment (*qua* emotional disposition), Prinz does not commit himself to the view that if an agent is not currently angry about rape, for example, then she does not currently judge rape to be wrong. As long as she retains a disposition to become angry when she hears about an instance of rape, that agent can be said to have an enduring moral judgement that rape is wrong.

(iii) Moral Sentiments and Realist Relativism

Having set out his analysis of sentiments, Prinz next provides an account of moral properties. He first suggests that:

> [a]n action has the property of being morally wrong just in case it causes feelings in the spectrum of both self-blame and other-blame emotions in normal observers under normal conditions.[79]

Prinz then rejects this proposal, for two reasons. Firstly, "normal conditions" are very hard to specify with any precision. Prinz suggests that the need to do so can be avoided altogether by appealing to the notion of a sentiment.

> [R]ather than saying that moral properties exist in virtue of causing certain emotions under certain conditions, we can say they exist in virtue of the fact that some observers have sentiments that dispose them to have those emotions.[80]

Secondly, the proposed account already contains an evaluative description of the obsever: "[t]he term 'normal observers' is troublesome because normality is an evaluative term".[81] Here,

---

[76] Prinz (2006) p. 34.
[77] Prinz (2007) p. 84.
[78] Prinz (2007) p. 85.
[79] Prinz (2007) p. 90.
[80] Prinz (2007) p. 92.

Prinz is concerned not to illicitly smuggle normative terms into an account of normativity itself. He argues that it is not possible to provide an objective account of who is to count as "normal", and that the solution is therefore to "define moral properties in terms of observers, and drop the normal bit".[82] It is not clear, however, that this move is in fact necessary. Prinz seems to conflate two different senses in which the term "normal" might be used. A normal observer might be one who is representative of a statistical majority of the population in terms of the moral judgements she makes. The use of this conception of normal in a definition of moral properties is certainly problematic. However, there is another sense in which normal might be used. Specifically, it could denote an observer who does not suffer from extreme epistemological impoverishment. To return to a topic touched upon earlier in the chapter, a psychopath might provide a good example of just such an epistemologically impoverished agent. If psychopaths are cognitively unable to experience certain types of emotional response, then they are not reliable sources of information with regard to the sorts of action that occasion such responses. This is not to discredit their opinions because they do not conform to those of the majority; rather, it is to make a principled distinction between agents who are psychologically able to perform a certain task (i.e. recognising moral properties), and those who are significantly less able to do so. In the latter case, it seems more reasonable to talk of "normal" observers than Prinz allows. This is a proposal which Prinz could in principle accommodate, given his brief, summary remark that a "sentiment represents the secondary quality of causing disapprobation in you (and others like you) under good epistemic conditions".[83]

Nevertheless, Prinz does in fact eschew the notion of a normal observer from his account. This leads him to reformulate his conception of moral properties as follows:

> [a]n action has the property of being morally wrong (right) just in case there is an observer who has a sentiment of disapprobation (approbation) toward it.[84]

The fact that Prinz does not make moral properties dependent upon the existence of a "normal" observer, who is responsive to those properties, gives his account some force against Joyce's argument for an error theory. Had Prinz drawn on the notion of a normal observer, Joyce would have been perfectly placed to object that *there can be no normal observers*. This is because the evolutionary process has deeply influenced how all of us think about morality. As a result, we cannot be trusted to form reliable beliefs about which moral properties actually exist. Prinz's position is not vulnerable to this objection, however. Because moral properties are generated by an observer's moral sentiments, the claim that those sentiments are influenced by evolution poses no threat to his account.

Clearly, Prinz's is a strongly relativistic position. He also argues, however, that it is also one which is compatible with metaethical realism. Prinz makes this claim as a result of his Humean, empiricist commitments. As discussed at the start of this chapter, Prinz sees his task as one of characterising and explicating ordinary moral psychology. Importantly, Prinz agrees with Joyce that pre-theoretical moral discourse is committed to realism. In his review of *The Evolution of Morality*,

---

[81] Prinz (2007) p. 92.
[82] Prinz (2007) p. 92.
[83] Prinz (2007) p. 102.
[84] Prinz (2007) p. 92.

Prinz notes that according to Joyce moral judgements, among other things, "aim to designate real properties".[85] Prinz goes on to say that he "basically agree[s] with this characterization".[86] Of course, accepting a realist account of ordinary moral discourse need not commit Prinz to metaethical realism; it is consistent with the explanatory aim of his project to give an account of *why* ordinary moral discourse is so committed, whilst simultaneously denying the truth of its realist pretentions. He could, that is to say, opt for an error-theoretic approach, as does Joyce. This is not the approach which Prinz adopts, however. Rather, he defends a form of causal realism.

According to Prinz's account, a property can be said to be real if it can be shown to possess causal efficacy; i.e. if it has the ability to bring about some change in the world. Prinz believes that moral facts qualify as real according to such an account. He writes that:

> [w]hen we say that stealing is wrong, we are not merely expressing our feelings, we are implicitly saying that stealing has the property of causing certain kinds of negative reactions in us. So the statement that stealing is wrong is true. By the same token, we can say that it is a fact that *stealing* causes these reactions in us, and thus the statement that stealing is wrong corresponds to a fact.[87]

So, Prinz denies that moral properties are objective in the sense traditionally argued for by moral philosophers, but he maintains that they are not therefore less real. Rather, moral properties exist as socially constructed facts. In this sense they are as objectively true as monetary facts: "[m]onetary value is created by us, but it is also a real feature of the world that has an impact on us".[88] Accordingly, Prinz characterises his account as a form of "constructive sentimentalism".[89]

It might be objected at this point that moral facts are importantly different from facts about the stock market: according to this thought, the former are emotionally constituted, whereas the latter are not. Given this difference, the objection continues, Prinz is not entitled to claim that moral facts admit of the same sort of objectivity as facts about the stock market. This line of objection is mistaken, however. According to Prinz's account, it is moral *judgements* which are emotionally constituted, not moral *facts*. The former are emotional responses, whereas the latter are those properties *to which* our moral judgements respond. An analogy can be drawn here between the emotion of fear, and the property of fearfulness. Suppose, in sympathy with Prinz's anti-nativist leanings, that we experience certain things as fearful only as a result of contingent facts about the process of our enculturation. It is simply true, according to Prinz's view, that those things towards of which we are afraid have the property of being fearful. How else are we to explain the fact that we react to them with fear? The fact that these things have the property of being fearful is not in any way made less true by the further fact that we were raised to believe this to be the case. How could the simple fact that I was raised in a culture which fears public humiliation, make public humiliation any less fearful for me? The fact that my fear is socially constructed does not undermine the truth of my claim that public humiliation is (for me) fearful. Thus, according to Prinz's view, moral facts can be the product of enculturation, and therefore relative, without being any less factual as a result.

---

[85] Prinz (2008c) p. 219.
[86] Prinz (2008c) p. 220.
[87] Prinz (2007) p. 166.
[88] Prinz (2007) p. 168.
[89] Prinz (2007) p. 167. Italics removed.

Prinz's approach has the advantage of not requiring of moral facts that they be intrinsically motivational. Moral motivation can be attendant upon the act of making a moral judgement, i.e. the act of entering a particular emotional state. His account is therefore broadly compatible with the response to Joyce's objection to naturalising moral properties which I outlined in Chapter Three (§7).

There is, however, more to Prinz's account than has so far been said. He does not hold that we are simply stuck with whatever moral dispositions we happen currently to have. Rather, our moral judgements are susceptible to a limited degree of rational correction. It is to this aspect of Prinz's theory which I now turn.

(iv) Moral Consensus and the Role of Reason

The previous section highlighted a tension between the relativistic and realist aspects of Prinz's constructive sentimentalism. The relativistic strand in this theory can presented even more starkly by highlighting Prinz's account of the plasticity of moral belief formation in children. Prinz holds that "morality emerges through the course of emotional conditioning".[90] He also holds that the results of such conditioning are more or less open-ended, and that "the wide range of moral rules found cross-culturally suggests that children can acquire moral attitudes toward just about anything".[91]

This position seems to sit uneasily with Prinz's further claim that the empirical study of moral discourse can generate new moral obligations, and that "moral progress is possible".[92] How does Prinz reconcile these views?

The answer to this question is that Prinz allows rational dialogue to play a role in the acquisition, rejection, and retention of moral attitudes. Suppose, for example, that we wish to convince an individual or community that some practice which they currently endorse is in fact immoral. It *might* be possible to do so by showing them that the practice in question falls under a particular description which, *by their own lights*, merits its disapprobation. In coming to see that the action in question can legitimately be so described, the individual or community with whom we are in dispute will also come to agree that their practice is, in fact, immoral. At this point they will (presumably) abandon it. Thus Prinz writes that:

> [s]ome rules are backed up by appeal to false factual knowledge. Opponents of women's suffrage claimed that women are too psychologically delicate for politics. This was probably a *post hoc* excuse for male dominance, but by resting the case against suffrage on a false factual claim, opponents opened up the door to moral revision through belief correction.[93]

---

[90] Prinz (2008c) p. 224.
[91] Prinz (2008a) p. 164.
[92] Prinz (2007) p. 289.
[93] Prinz (2007) p. 291.

It is this sort of process which Prinz has in mind when he writes that "we can figure out what our obligations are by figuring out what our moral beliefs commit us to".[94] More precisely, suppose Prinz is right that moral anger is a response to those situations which constitute a violation of someone's rights. When they recognised certain of their beliefs about the nature of women's psychology as false, the opponents of women's suffrage also came to see that denying suffrage to women violated their rights. Consequently, they came to feel moral anger towards that practice.

The process of moral belief revision described above has its limitations, however. At any given point in an ethical dispute, the disputants may find themselves disagreeing over one of their basic values (or "grounding norms", as Prinz sometimes calls them). As we saw earlier, when this happens there is nothing more to be said; dialogue effectively breaks down.

> When two people disagree about whether something is wrong, they can have a rational debate about whether it falls under a category about which they have a grounding norm. [...] But when it comes to grounding norms, rational debate is impossible.[95]

Thus, Prinz argues that:

> [if] the concept OUGHT is linked to the idea of [interpersonal] normative authority, then it seems to follow that missionaries are mistaken when they say that the Akamara ought to refrain from cannibalism, even if they are right to say that [according to the values of the missionaries] cannibalism is wrong.[96]

That is, in the absence of a shared grounding norm pertaining to the acceptability of consuming human flesh (or of showing due respect to the dead, or some other relevant norm), there is no potential for rational moral debate, and so no way of assessing opposing moral claims. It is the existence of shared grounding norms that generates our moral commitments to one another, and allows us to evaluate our actions as falling short of a standard to which we ourselves subscribe.

5. Objections to Prinz

Having provided an overview of Prinz's constructive sentimentalism, and of its relation to his philosophy of emotion, in this section I consider some of the objections which can be made against his account. I begin in §5.i with an objection made by Joyce, according to which Prinz's account fails to pick out any moral properties by failing to give an adequate account of what it is to possess an emotional disposition. I argue that Joyce's objection fails because it does not pay close enough attention to Prinz's theory of emotion. The second objection, which I consider in §5.ii, stems from more closely attending to the relation of Prinz's theory of emotion to his moral philosophy. In it, I argue that Prinz's account is internally inconsistent, owing to a tension between the claim that emotions *qua* appraisals must be capable of misrepresentation, and the idea that our basic values are not susceptible to rational correction. Finally, in §5.iii I discuss the different conceptions of moral

[94] Prinz (2007) p. 1.
[95] Prinz (2007) p. 125.
[96] Prinz (2007) p. 179.

realism in the literature on sentimentalism, and argue that Prinz's account fails to satisfy D'Arms and Jacobsen's criteria for a realist theory of morality.

(i) Prinz as Error Theorist

Joyce has objected that, as it stands, Prinz's sentimentalism fails to pick out any moral properties whatsoever. For this reason, he describes Prinz as "unwittingly offering an error theory".[97] Joyce's objection stems from what he sees as shortcomings in Prinz's account (or rather, the lack thereof) of the circumstances in which agents can be said to possess various dispositions. He argues that:

> in failing to specify [an agent's] circumstances Prinz has provided a description of the disposition that is incomplete to such an extent that it fails to denote any property at all [...].[98]

Joyce fleshes this objection out by asking:

> does *anyone* have the disposition to feel the occurrent emotion of bitterness *period*? Do you? The natural question is "At what?" But even if "At what?" could be answered – suppose it's specified that we're asking whether you have the disposition to feel bitterness towards ex-lovers – the next question is "In what circumstances?" You might feel occurrent bitterness towards ex-lovers in certain circumstances, but not in other circumstances. [...] Prinz's description of the sentiment of resentment, as it stands, fails to denote any property at all in any possible world, and thus, if we take the definition at face value, he has offered an error theory of resentment. And since he has tied moral properties to these sentiments, [...] he has also offered a moral error theory.[99]

Joyce argues that the problem is not simply that Prinz's description of an emotional disposition is imprecise; it is rather that this description is in fact "incomplete in a striking manner which leaves [Joyce] with *no idea* of how it should be finished".[100]

It is possible to question whether this account is as incomplete as Joyce takes it to be. Prinz's philosophy of emotion contains a detailed discussion of the higher cognitive emotions (of which the moral emotions are a species). As we have seen, these are characterized as representing concerns: the organism-environment relations that pertain to an organism's well-being. Now, Joyce has (relatively) little complaint with Prinz's claim that a fear of flying "is a standing state that will become an occurrent experience when, for example, you board an airplane".[101] It is not obvious, therefore, why the analogous claim that moral anger is a disposition that will become an occurrent state when

---

[97] Joyce (2011) p. 163.
[98] Joyce (2011) p. 162.
[99] Joyce (2011) p. 163.
[100] Joyce (2011) p. 163.
[101] Prinz (2007) p. 85.

a human rights violation is perceived should be any more problematic than the case of a fear of flying.

This line of reply emphasises the importance of reading Prinz's metaethical account as an extension of his philosophy of emotion. This is something which Joyce's objection fails to appreciate. A contiguous reading of these theories (as undertaken in this chapter), however, highlights apparent discontinuities between the two which create problems for the internal consistency of Prinz's account.

(ii) An Internal Inconsistency in Prinz's Account

It will be recalled that an essential feature of a representation is that it be capable of *mis*representation. To what extent, then, do moral emotions possess this potential?

Prinz's answer to this question seems ambivalent. In his causal account of moral properties, he compares them with other, non-moral emotion inducing properties, such as disgustingness. He notes that:

> [t]here are situations in which one has an emotion without the emotion-inducing property. Suppose, for example, that I form the false belief that a jar of food in my refrigerator has spoiled. I might discard it because I am disgusted, or because I believe it is disgusting, but not because it actually is disgusting. [...] But now consider the case where the food actually is spoiled. In this case, my disgust is caused by the fact that the food is disgusting. The emotion-inducing property is causally responsible for my mental state, and, thus, is implicated in my subsequent behavior.[102]

This certainly suggests that moral emotions can be mistaken in relevantly similar ways to non-moral emotions. The phenomenon of moral misrepresentation is exemplified by Prinz's discussion of changing attitudes towards women's suffrage. Here, a practice is reconceptualised as falling under a particular description which was previously thought not to apply to it (i.e. a violation of human rights).

However, Prinz's account also contains a commitment to the idea that some aspects of an agent's moral psychology are not capable of being mistaken: namely, her basic values. These are beyond rational discussion, and, importantly, are taken to be emotionally constituted in the same way as other moral emotions: "[w]hen we get down to basic values, passions rule".[103] If basic values are rationally non-negotiable, then it would seem to follow that one simply cannot be mistaken in subscribing to them, whatever they may be.

There is therefore an important question to ask of Prinz's account: are basic values incapable of misrepresentation? This question generates a dilemma for Prinz. Opting for the first horn of the dilemma, Prinz could assert that basic values *are* incapable of misrepresentation. However, it would

---

[102] Prinz (2007) pp. 166 – 167.
[103] Prinz (2007) p. 32.

then follow that according to his account, basic values, unlike other emotions, are not representations. In the absence of a principled reason for taking basic values to be somehow different from all other emotions, this generates the possibility that *other* emotions may not be representations either. Such a conclusion would pose a serious challenge to Prinz's embodied appraisal thesis, and is therefore one which he would wish to avoid. The second horn of the dilemma looks equally unappealing, however. If basic values *are* supposed to be capable of misrepresentation, then it becomes unclear how they can perform their function as fundamental emotional states, the sharing of which makes inter-personal moral debate possible. This is because each basic value can be seen to stand in need of further justification, in order to ensure that it is not a misrepresentation. Such justification could only be cognitive in form, as any purported non-cognitive justification would simply reiterate the dilemma. A cognitive justification of basic values would effectively amount to the abandonment of non-cognitivism, however. It is therefore a result which Prinz would find extremely unappealing. Thus, whichever horn of the dilemma Prinz opts for will create serious problems for the coherence of his overall project.

(iii) Can Prinz Project His Cake and Eat it Too?

Prinz's formulation of constructive sentimentalism is designed to allow him to avoid an objection which is often levelled at sentimentalist theories of ethics. This objection claims that sentimentalism is unable to provide a satisfactory account of moral realism because it must choose between two rival, yet equally flawed, metaphors in its account of moral epistemology: those of perception and projection.

D'Arms and Jacobson have discussed the shortcomings of both of these metaphors. They argue that the metaphor of perception "exaggerates the similarity between the sentiments with which we 'perceive' values and ordinary faculties of perception".[104] Although it captures the apparently objective nature of moral judgement, this exaggeration overstates the extent to which values can be described as existing independently of valuing agents. Whilst the metaphor of projection is not vulnerable to this sort of objection, it struggles to portray morality as truly objective. That is, "[b]y encouraging an understanding of values as figments of our feelings, it threatens to undermine the practice of moralizing".[105]

Prinz hopes to avoid the problems associated with each of these metaphors by building *both* of them into his account of constructive sentimentalism. Accordingly, he argues that:

> [t]here is a sense in which we perceive moral properties. In fact, one can make literal sense of this idea by noting that emotions are perceptions of changes in the body, and it is by means of our emotions that we come to discover that something is right or something is wrong. But there is also something apt about the idea of moral projection.

[104] D'Arms and Jacobson (2006) p. 212.
[105] D'Arms and Jacobson (2006) p. 212.

> Morals are not objective features of the world in the way that, say, lions and tigers and bears might be. They come from us.[106]

Is Prinz right to think that constructive sentimentalism avoids the problems associated with the metaphors of perception and projection, or does it in fact inherit both sets of concerns? The answer to this question will depend to some extent on how the sentimentalist project is conceived. It was suggested in §2.ii that the disagreement between the conception of sentimentalism endorsed by D'Arms and Jacobson and that endorsed by Prinz may not be as deeply conflicting as appearances first suggest. It is now time to explore this suggestion more fully.

As previously discussed, D'Arms and Jacobson hold that sentimentalism should "preserve the idea that values are somehow grounded in the sentiments, while at the same time making sense of the rational aspects of evaluation".[107] Prinz takes himself to have discharged this obligation. His version of sentimentalism allows for the moral evaluation of actions according to whether or not they are consistent with (i.e. are accurately represented as extensions of) the basic values to which an individual or community subscribes.

As a result of the primacy of the basic values, Prinz's account can be described as one in which moral reasons "come from within ethics; that is, these reasons will themselves entail substantial ethical commitments".[108] This, however, is a position which D'Arms and Jacobson find "deeply problematic".[109] The problem is that, in order to show that a newly revised set of moral values is normatively better than its predecessor, it is insufficient merely to show that those values have greater internal consistency:

> until they distinguish the relevant notion of appropriateness, [sentimentalists have not] yet "earned the notion of truth" –or, if you prefer, of knowledge or objectivity – for evaluative judgment.[110]

Has Prinz met this challenge?

The answer to this question turns on the notion of "objectivity" in play. Prinz rejects mind-independent accounts of moral objectivity,[111] whilst insisting that his version of sentimentalism makes moral judgements truth-apt. Certainly, D'Arms and Jacobson do not mean to suggest that sentimentalism must provide anything like a mind-independent account of moral objectivity: this would be to abandon sentimentalism's "modest, anthropocentric conception of value"[112] altogether. Rather, they propose a "rational sentimentalism",[113] according to which sentiments be considered appropriate, or "fitting".[114] A sentiment is fitting when it is possible:

---

[106] Prinz (2007) p. 168.
[107] D'Arms and Jacobson (2000) p. 722.
[108] D'Arms and Jacobson (2000) p. 733.
[109] D'Arms and Jacobson (2000) p. 733.
[110] D'Arms and Jacobson (2000) p. 736.
[111] Prinz (2007) chapter 4.
[112] D'Arms and Jacobson (2006) p. 212.
[113] D'Arms and Jacobsen (2000) p. 746.
[114] D'Arms and Jacobsen (2000) p. 746.

> [t]o endorse the response in the relevant way, which constitutes taking the circumstances to be genuinely ϕ. [...] [I]n order to provide any substantive grounding for the distinction between reasons that are and are not relevant to the fittingness of an emotion, it is necessary to examine our actual emotions piecemeal, in order to articulate the differences in how each emotion presents some feature of the world to us when we are in its grip.[115]

Prinz's account achieves this in relation to the higher cognitive moral emotions, but denies that such a program can bear fruit when applied to basic values. D'Arms and Jacobson would therefore deny that his account has earned the right to talk of moral truth. Is there any way to resolve this dispute which does not reduce to a clash of intuitions about what constitutes realism?

The most promising resource with which to meet D'Arms and Jacobson's challenge can be found in further unpacking Prinz's account of basic values. This may seem surprising, as it is also this account which creates the tension between the two.

D'Arms and Jacobson's concern is that a non-rationalist sentimentalism offers no principled way of deciding which sentiments to endorse and which to revise when attempting to make an extent set of such sentiments more internally consistent. This concern construes each such sentiment as being equally contingently endorsed, and therefore as equally potentially revisable. Prinz's account of basic values rejects this assumption of equipotential value revision. According to Prinz, there are some values which we simply *cannot* decide no longer to endorse: they are a basic fact of our psychological or cultural identity. As these values are both non-volitional and non-revisable, they can provide the foundation for a principled way of evaluating non-basic moral emotions. Indeed, this is precisely the role which Prinz sees them playing.

Is such a response ultimately satisfactory? It certainly avoids the charge levelled against the metaphor of projection: i.e. that it does not accurately represent the feeling of objectivity attendant upon moral judgements. If basic values are a brute, non-volitional fact about an agent's psychology, then they may be experienced as being just as objectively true as any other such fact.

Despite this, however, Prinz's view is unlikely to be considered attractive by many moral realists. It is generally considered to be a part of a realist account of morality that it be capable of offering normative guidance to those faced with a moral dilemma. Prinz's sentimentalism *can* do this, but only *some* of the time. Moreover, his commitment to a robust form of relativism means that *when* it can do so will be largely a matter of luck (depending on whether the disputants share the relevant basic values), rather than due to the specific features of the dilemma in question. Given this restriction in the scope of its normative role, Prinz's theory does not reflect one of the archetypal features of moral realism. It is realist by name, but not by nature.

6. Conclusion

[115] D'Arms and Jacobson (2000) p. 746.

This chapter has critically assessed Prinz's theory of realist sentimentalism. Prinz takes moral judgements to be non-cognitive, emotionally constituted appraisals of organism-environment welfare. The formation of specific moral judgments is dependent upon social conditioning, and morality is therefore importantly relativistic. Nevertheless, Prinz argues that moral judgements are responses to real moral properties. He does this by endorsing a form of causal realism, according to which moral properties are to be identified simply as those properties which cause us to experience a particular moral emotion.

I have argued that Prinz's theory fails on two accounts. Firstly, Prinz is unable consistently to maintain both that basic values are not susceptible to rational assessment, *and* that, *qua* emotions, basic values are embodied appraisals. If basic values are embodied appraisals then they must be capable of misrepresentation, yet to admit this is to thereby deny that they cannot be rationally critiqued.

Secondly, Prinz's account does not satisfy the criteria for moral realism as specified by D'Arms and Jacobsen. Whilst Prinz's theory *is* able to offer a naturalistic account of moral properties, his account is not one which can reliably be used to settle normative debates. For a metaethical perspective to be satisfactorily realist, it must be more amenable to non-relativistic first-order moral discussion.

Of course, if Prinz's theory is *as* realist as naturalistic metaethics can legitimately get, then the complaints which I have made against it might be thought to carry less weight. After all, should not the moral realist be satisfied with however much realism she can get? In the next, and final, chapter, I will defend what I take to be a more robustly realistic form of sentimentalism. My approach will be a care-ethical one, and will draw not only on some of the insights in Prinz's account (albeit supplemented by evolutionary considerations), but also on the work of Haidt (discussed in Chapter Four). Furthermore, it will emphasise the importance of the role played by cultural evolution (discussed in Chapter One) in articulating a realist theory of morality. An evolutionary care-ethics, I will argue, is the best perspective from which to counter Joyce's argument for an error theory.

Chapter Six

Empathetic Caring as the Foundation of Moral Judgement

1. Introduction

In this final chapter I defend an evolutionary approach to care-ethics as the most promising realist alternative to Richard Joyce's evolutionary error theory. The account which I defend sees moral judgements as based on our capacity to ascribe intentional psychological states to other agents. Essentially, moral evaluation is to be understood as an attempt to determine whether or not an action was motivated by empathetic concern for the wellbeing of another. Moral judgements, on this view, are true or false depending on whether or not they are the product of an accurate representation of an agent's intentional state.

In §2 I give a short summary of what care-ethics is, of its origin in feminist developmental psychology, and of how it differs from other moral theories. §3 then provides a more detailed analysis of the notion of care. I discuss the close connection between the notions of care and empathy, and show why, despite this close connection, it is not possible to talk exclusively about empathy rather than care.

With a sufficiently precise conception of caring having been established, I go on to argue, in §§4 – 5, that there are excellent reasons to think that our capacity for empathetic caring forms the foundation of our moral judgements. I begin, in §4, by noting the distinction drawn in moral psychology between operative and expressed moral principles. There, I argue that our expressed moral principles can be interpreted as culturally evolved heuristics, and that these heuristics are subordinate to universal, operative moral principles. These operative principles are responsible for our moral intuitions, and we use them to assess the truth or falsity of suggested expressed principles. I then, in §5, provide good evolutionary evidence in favour of interpreting our operative moral principles care-ethically. I discuss research which suggests that social selection pressures, which were made crucially important when our ancestors became ecologically dominant, underlie the evolution of our capacity for theory of mind and empathetic perspective-taking. I show how this evidence complements the moral-psychological data described in §4, and how it supports my claim that our operative moral intuitions ought to be interpreted care-ethically.

With an evolutionary metaethics of care having been defended, in §6 I give a short argument in favour of a normative ethics of care. I ask whether the possible ecological rationality of moral heuristics counts against adopting a normative care-ethics, and argue that this is not the case. Firstly, identifying caring agency as the target attribute of moral heuristics allows us to minimise the risk generated by thinking about moral dilemmas without relying on those heuristics. Secondly, moral heuristics need not be altogether abandoned; they will still serve a purpose much of the time, but recognising their normative limitations permits us to dispense with them in situations in which they fail to track their target attribute.

In §7, I argue that an evolutionary care-ethic is a realist theory of morality. I defend this claim by considering reference fixing accounts of moral realism endorsed by Michael Slote and Joshua Gert, and by arguing that the metaethical perspective outlined in this chapter should be understood in reference fixing terms. I then show that my account is better able to meet the criticisms made against Jesse Prinz's brand of realist sentimentalism in the previous chapter, as evolutionary data allows Prinz's notion of basic values to be cashed-out non-relativistically. I also argue that my account is an improvement on Slote's, as it fixes the reference of moral terms by appealing to their Darwinian function, rather than by appealing to contingent facts about our sentimental dispositions.

Finally, in §8 I briefly recap how the metaethical perspective defended by this chapter constitutes a reply to Joyce's argument for an evolutionary error theory. I show how an evolutionary care-ethics can accommodate many of Joyce's claims about the nature of moral judgement and the non-negotiable commitments of ordinary moral discourse. Unlike Joyce's evolutionary aetiology, however, my account of the evolution of morality *does* give us reason to believe in moral properties, *qua* facts about agents' intentional states. The theory which I defend therefore constitutes an evolutionary success theory, and Joyce's argument for an error theory is refuted.

2. Care-Ethics: An Overview

Care-ethics is a relatively new approach to moral philosophy, finding its first clear statement in 1984 with Nel Noddings' *Caring: A Feminine Approach to Ethics and Moral Education*. However, the seeds from which Noddings' work grew had been sown two years earlier, by Carol Gilligan's *In a Different Voice: Psychological Theory and Women's Development*.

Gilligan took issue with a model of moral development which had previously been proposed by developmental psychologist Lawrence Kohlberg. She argued that this model illegitimately represented women as being "deficient in moral judgment".[1] Kohlberg constructed a six-stage model of moral development. According to this model, the highest stages of development are achieved when "relationships are subordinated to rules (stage four) and rules to universal principles of justice (stages five and six)".[2] Kohlberg found that women typically did not attain such levels, however. Rather, their development tended to stop at around stage three, at which "morality is conceived in interpersonal terms and goodness is equated with helping and pleasing others".[3]

According to Gilligan, this putative deficiency in women's moral judgement was actually nothing of the sort. Rather, she argued that Kohlberg's findings reflected a deficiency in his research, such deficiency being the product of Kohlberg's having used an all-male group of subjects in his case studies. Gilligan argued that this exclusive focus on a male perspective delegitimised the universal applicability which Kohlberg claimed for his account. Rather, the third stage of development which he identified represented an alternative, typically feminine way of conceptualising moral problems. According to this feminine perspective, a moral dilemma arises as a result of:

---

[1] Gilligan (1993) p. 18.
[2] Gilligan (1993) p. 18.
[3] Gilligan (1993) p. 18.

> conflicting responsibilities rather than from competing rights and requires for its resolution a mode of thinking that is contextual and narrative rather than formal and abstract.[4]

Kohlberg was mistaken to see deficiency in the moral reasoning of women, Gilligan argued. As an alternative view, she suggested that: "men and women may speak different languages that they assume are the same, using similar words to encode disparate experiences of self and social relationships".[5] Thus, Gilligan continues:

> in the different voice of women lies the truth of an ethic of care, the tie between relationship and responsibility, and the origins of aggression in the failure of connection.[6]

In *Caring*, Noddings builds upon Gilligan's research to develop a contextualist critique of traditional approaches to moral philosophy. She criticises the history of western ethics as being conducted "largely in the language of the father: in principles and propositions, in terms such as justification, fairness, justice".[7] In contrast to this tradition, Noddings argues that "the memory of caring and being cared for [...] form the foundation of ethical response".[8] For Noddings, such caring is to be characterised as a motivational disposition, one in which:

> caring is directed toward the welfare, protection, or enhancement of the cared-for. When we care, we should, ideally, be able to present reasons for our action/inaction which would persuade a reasonable, disinterested observer that we have acted in behalf of the cared-for.[9]

Consequently, morality is seen by Noddings as a "commitment to behave in a fashion compatible with caring,"[10] such commitment being "rooted somehow in common human needs, feelings, and cognitions".[11]

Central to Noddings' conception of morality was the claim that moral judgement must be contextual. It is true that, for the care-ethicist, certain actions will very regularly be deemed morally unacceptable. However, the temptation to introduce universally applicable moral principles in order to account for this regularity must be avoided. "We do not say: It is wrong to steal. Rather, we consider why it was wrong or may be wrong in this case to steal".[12]

Work on care-ethics initially took place primarily within the confines of feminist philosophy, with the exception of developmental psychologists such as Gilligan. Philosophical opinion was divided between those who saw care-ethics as an exciting opportunity to develop a specifically feminist branch of moral philosophy, and those who saw in it a reinforcement of gender stereotypes.

---

[4] Gilligan (1993) p. 18.
[5] Gilligan (1993) p. 173.
[6] Gilligan (1993) p. 173.
[7] Noddings (2003) p. 1.
[8] Noddings (2003) p. 1.
[9] Noddings (2003) p. 23.
[10] Noddings (2003) p. 91.
[11] Noddings (2003) p. 27.
[12] Noddings (2003) p. 93.

Patricia Scaltsas worried, for example, that "[i]f the virtues of empathy and care are [...] sex determined, then women are trapped biologically in a restrictive, feminine stereotype."[13]

Concerns of this sort encouraged a close critical examination of Gilligan's research. In the light of that examination, the apparent correlation between moral perspective and sex was found to be less evident than Gilligan's results suggested. As Carol Tavris reports:

> when subsequent research directly compared men's and women's reasoning about moral dilemmas, Gilligan's ideas have rarely been supported. In study after study, men and women use *both* care-based reasoning [...] *and* justice-based reasoning [...]. In study after study, researchers report no average differences in the kind of moral reasoning that men and women apply.[14]

This is not to say, however, that the notion of an ethic of care should itself be abandoned. As Tavris herself notes, the findings which she reports in fact "confirm Gilligan's argument that people make moral decisions not only according to abstract principles of justice, but also according to principles of compassion and care".[15] Gilligan's identification of a moral perspective different from that of traditional, justice-based approaches therefore remains important. If her association of the care perspective with women was an error, then the correction of this error now allows care ethicists to argue that their conclusions apply with equal validity to *either* sex. By rejecting essentialist interpretations of care-ethics, then, it is possible to see in such an approach the potential for "nothing less than a total or systematic *human* morality".[16]

This chapter will argue that caring does in fact form the foundation of human moral judgement. Before making this argument, however, I first need to give a somewhat fuller account of the notion of caring. This is the task of the next section.

3. Caring

As we have just seen, Noddings takes caring to be an affective attitude. To care is to have a concern for the "welfare, protection, or enhancement"[17] of the one towards whom that care is directed. Noddings glosses this by saying that if we care about another, then what that other "think[s], feel[s], and desire[s] will matter"[18] to us.

The psychology of the caring agent has received a somewhat more detailed treatment in the work of Michael Slote. Slote argues that the concept of caring should pick out a disposition towards empathetic concern on the part of the caring agent. He notes that Noddings has expressed scepticism about conceptualising caring in such a way, arguing that the phenomenology of caring is

---

[13] Scaltsas (1992) p. 20.
[14] Tavris (1992) p. 85.
[15] Tavris (1992) p. 85.
[16] Slote (2007) p. 3.
[17] Noddings (2003) p. 23.
[18] Noddings (2003) p. 9.

distinct from that of empathising. Slote avoids this objection by arguing that Noddings' criticisms were directed towards *projective* empathy, whereas he endorses *mediated associative* empathy. Projective empathy occurs when we actively imagine ourselves to be in the situation of another, drawing upon our own beliefs about what is good, desirable, or of value in the process. Mediated associative empathy, Slote contends, is far closer to the psychological state which Noddings envisages in terms of caring. This is a state in which the empathiser "pay[s] attention to, and [is] absorbed in, the way the other person structures the world and his or her relationship to the world".[19] Thus, for Slote, there is no conflict between the notions of empathy and of caring. Indeed, he argues that:

> distinctions of empathy and of *empathic* caring correspond better to common-sense moral distinctions than does anything that can be understood by reference to caring taken, so to speak, on its own.[20]

Why should this be the case? Simply put, Slote holds that differences in the degree to which we are able to empathise with others will generate differences in our intuitions about the moral responsibilities we have towards those others. He elaborates upon this idea in a discussion of abortion. For Slote, we are better able to imaginatively empathise with late-stage foetuses, which are in many respects similar to neonates, than we are early-stage foetuses. The latter "look more like fish or salamanders [...] and they lack experience, a brain, and even limbs".[21] Slote uses this difference in potential for empathetic engagement to account for (and endorse) the intuition that early-stage abortion is not morally problematic, whereas late-stage abortion is. I will not critically assess this claim here, deferring such discussion until §7.ii. At present it is sufficient to observe that incorporating an empathetic element into the notion of care allows Slote to account more fully for differences in our moral evaluations.

Slote summarises his position by stating that "empathy is a crucial source and sustainer of altruistic concern or caring about (the wellbeing of) others".[22] Thus, he argues that:

> [t]he ethics of empathic caring evaluates the actions of individuals in terms of whether they express, exhibit, or reflect empathically caring motivation, or its opposite, on the part of individuals.[23]

This characterisation raises this question of whether it is possible to replace talk of caring with talk of empathising entirely. This is a *prima facie* attractive proposal, as it would remove much of the conceptual imprecision which comes with talk of care. Such a move would be too quick, however. Empathy, as Slote uses the term, "involves having the feeling of another (involuntarily) aroused in ourselves".[24] Slote distinguishes this from sympathy, in which we are aware of, for example, the suffering of another, "and positively wish them well".[25] This is an importantly different concept: it does not involve a shared *experience* of suffering, but an *awareness* of suffering in

[19] Slote (2007) p. 12.
[20] Slote (2007) p. 14.
[21] Slote (2007) p. 18.
[22] Slote (2007) p. 15.
[23] Slote (2007) p. 94.
[24] Slote (2007) p. 13.
[25] Slote (2007) p. 13.

another. Furthermore, such awareness comes with the desire to alleviate the suffering of which we are aware. This desire did not form a part of Slote's definition of empathy, which is restricted to the involuntary arousal in oneself of the feelings perceived in another. To capture both shared experience of feelings *and* a desire to render aid, then, we need the notion of care. By eschewing talk of caring in favour of talk about empathy, that is to say, we risk losing the connection between an agent's possessing a morally relevant affect, and that agent's having a concern for the welfare of the person with whom he or she empathises.

In addition to this concern, it is also possible that an emphasis on empathy rather than care would obscure the connection between caring and action. Virginia Held has argued that, whilst possession of the appropriate affective attitude is an essential feature of the notion of care, there is insufficient emphasis in traditional accounts "on actually meeting needs".[26] That is, we must understand care not only in terms of an agent's intentionality, but also in terms of how that agent acts with regard to the object of his or her care. Caring, for Held, "shows us how to respond to needs and why we should".[27] Nevertheless, this practical aspect of caring cannot be understood in isolation from the notion of caring as a value: "we need care as a value to pick out the appropriate cluster of moral considerations, such as sensitivity, trust, and mutual concern, with which to evaluate such practices".[28]

Although the psychology of the empathetic agent can help us to make sense of why these moral considerations prove salient, an exclusive focus on empathy could lead us to pay insufficient attention to the importance of how agents *act* on those considerations. For the purposes of this discussion, then, attention will remain focussed on caring as such.

The foregoing comments on care do not provide a precise definition of what caring is, and of when an agent is engaged in it. Whilst a lack of specificity on this issue may seem to undermine the potential to engage in a rigorous discussion of the concept, to attempt to give a thorough definition would be to ignore care-ethics' commitment to contextualism. There can be no abstract formulation of care-ethical principles; rather, concrete situations must be described with as much detail as possible. Only then will it be possible to establish what constitutes a caring response, given the particular facts of the matter at hand.

In addition to the contextualism of care-ethics, there is a second reason for thinking that attempting to precisely formulate the concept of care would prove problematic. This line of thought focusses on the role which considerations of care play in our moral psychology. If care-ethical intuitions generate operative rather than expressed moral principles, they may for that very reason be much harder to formalise. More will be said about the distinction between operative and expressed principles in the next section.

So far, I have introduced and outlined the concept of caring. I have discussed psychological research which suggests that both men and women can adopt a caring perspective when assessing certain moral dilemmas, and I have noted that this perspective has been endorsed by some feminist philosophers as the most appropriate way to engage with any moral issue. I have not yet provided

[26] Held (2006) p. 20.
[27] Held (2006) p. 42.
[28] Held (2006) p. 38.

any argument that care-ethical intuitions form the foundations of our moral judgements. Nor have I argued that a care-ethical approach is *in fact* the most reliable way of thinking about morality. I make these arguments over the course of the next three sections. I begin by discussing recent research in moral psychology, describing how the notion of care fits into our current understanding of how we make moral judgements. I then support this interpretation by appealing to research into the evolution of social intelligence and theory of mind in modern humans. This research offers significant support for the thesis that our moral judgements are structured around caring. Having identified care-ethics as the basis of our moral intuitions, I then consider whether we should consequently adopt a normative ethics of care.

4. Care-Ethics and Moral Psychology

Earlier chapters have already discussed the phenomenon of moral dumbfounding, as presented by Jonathan Haidt.[29] In the previous chapter, moral dumbfounding was seen to be an important piece of empirical evidence in support of Prinz's argument for the emotional constitution of morality. The recalcitrant moral intuitions which Haidt draws attention to are, Prinz argues, "basic values": emotionally constituted conversation stoppers, the moral force of which is not subject to the influence of reasoned debate. However, the phenomenon of moral dumbfounding also points to features of our moral psychology which lie beyond the issue of how moral judgements are constituted.

(i) The Importance of Intentionality to Moral Judgment

Hauser et al (2007) conducted a study designed to determine whether subjects' moral judgements were influenced by considerations of double effect. That is, they explored

> how the psychology of such distinctions as that between killing and letting die and intended harm and foreseen harm bears on the nature of our moral judgments.[30]

There were two aspects to this research, both of which are relevant to the conclusion argued for in this chapter. The first aspect was to explore the "dissociation between judgment and justification for certain moral dilemmas"[31] identified by Haidt, i.e. the phenomenon of moral dumbfounding. The second aspect of Hauser et al's research was "to show how our moral judgments are mediated by an appraisal system that takes into account the causal and intentional properties of human action".[32]

The authors used a variety of moral dilemmas to test subjects' moral intuitions across a range of scenarios. The situations described in these dilemmas were specified in ways which allowed

[29] In Haidt (2001).
[30] Hauser et al (2007) p. 2.
[31] Hauser et al (2007) p. 2.
[32] Hauser et al (2007) p. 2.

the authors to isolate which particular features of the dilemma proved morally salient to subjects' intuitions. Thus, for example, it was possible to specify dilemmas which differed only according to whether a specific harm was intended or merely foreseen by the agent facing the dilemma.

It was found that subjects' judgements *were* affected by the representation of an agent's intentional attitude, and that these judgements did not vary across subpopulations of test subjects. The authors concluded that "across a variety of nationalities, ethnicities, religions, ages, educational backgrounds (including exposure to moral philosophy), and both genders, shared [moral] principles exist".[33] Interestingly, despite the apparent universality of these shared moral principles:

> a large majority of subjects failed to sufficiently justify their moral judgments, including a majority of those subjects who had been exposed to readings in moral philosophy.[34]

This observation led the authors to the conclusion that it is possible to distinguish between "operative" and "expressed" moral principles. The latter are those moral rules of thumb to which we have conscious access ("do not lie", "do not steal", etc.), whereas the former are intuition generating principles which are unavailable to conscious introspection. According to Hauser et al, then, in certain conditions:

> intuition drives subjects' judgments, and with little or no conscious access to the principles which distinguish between particular moral dilemmas. [...] [Thus] when people make certain kinds of moral judgments, they may do so without consciously applying explicitly understood principles.[35]

These unconscious principles seem particularly sensitive to "the causes and consequences of action, especially its intentional structure".[36] As Hauser et al observe, "[p]revious work in moral development, especially as championed by Kohlberg and his students, failed to make these distinctions, focussing exclusively on expressed principles".[37] As a result, the importance of intentionality in generating our moral intuitions was largely overlooked.

Hauser et al.'s findings complement Haidt's work on moral dumbfounding, suggesting that this phenomenon occurs when non-conscious, "operative" moral principles generate moral intuitions. Further, these operative moral principles are likely to be regulated by (perceived) facts about the intentional characteristics of agency.

The structure of these operative principles suggests that they are sensitive to the sort of considerations which care-ethicists see as central to moral judgement. However, more needs to be said before concluding that operative principles are concerned *specifically* with caring intentions. I take up this issue in the next section. Before embarking upon that discussion, however, it is necessary to say more about the relationship which holds between operative and expressed moral principles.

---

[33] Hauser et al (2007) p. 16.
[34] Hauser at al (2007) p. 16.
[35] Hauser et al (2007) p. 17.
[36] Hauser et al (2007) p. 17.
[37] Hauser et al (2007) p. 2.

(ii) Operative and Expressed Moral Principles

As we have seen, care-ethicists hold that the moral status of an action always follows from intentional facts about the psychology of the agent who performed it. They do *not* hold that the moral status of an action follows only *sometimes* from facts about intentionality, and *sometimes* from facts about whether an action conforms to a general, universally applicable principle. Yet the research by Hauser et al explicitly attests that operative principles exist alongside expressed ones. This is a problem for care-ethics, as it suggests that moral judgement is *not* based exclusively on considerations of care.

The simplest way of dealing with the existence of expressed principles is to deny their normative weight. Accordingly, whilst we may all sometimes use such principles in deciding how to act, or in evaluating the actions of others, we ought not to do so. Because they are not sensitive to considerations of care, such principles are likely to provide misleading guidance. Such a position has in fact been taken up by some care-ethicists. For example, Slote (2007) argues that "it makes sense to think of the/an ethics of care as covering all, and not just some smaller part, of morality".[38] Unsurprisingly, opinion is divided on this issue. Thus, Held (2006) argues that "neither justice nor care can be dispensed with: Both are extremely important for morality".[39]

Typically, the success of either approach has been thought to depend on its ability to provide answers to problems which the other is unable to resolve. Here I develop a less direct, though in my view more appealing, line of argument in favour of an exclusively care-ethical perspective.

According to my proposal, our introspectively available expressed principles (such as those found in utilitarian and deontological moral theories) are themselves governed by specifically care-ethical operative principles. Expressed principles are culturally-evolved rules-of-thumb: deliberative shortcuts which recommend certain actions as compatible with a caring attitude, or prohibit certain others as incompatible with such caring.[40] As a result of their generality, however, our expressed moral principles will, in certain atypical circumstances, fail to recommend a caring course of action. In these exceptional circumstances, our care-ethical operative principles step in to veto the course of action which the expressed principles pre- or proscribe.

This interpretation is of course speculative, but there are good reasons to take it seriously. To begin with, it allows us to make sense of the fact that we typically regulate our expressed principles according to our (operative) moral intuitions. This can be seen by considering some of the traditional deontological objections to act-utilitarianism. Deontologists argue that in certain circumstances, acting so as to bring about the greatest amount of (for example) happiness will entail acting in ways which violate our duties towards others. Thus it might be the case that in order to maximise utility, one is required to frame an innocent person for a crime she did not commit, murder a healthy hospital patient in order to use her organs to save the lives of five others, or torture an innocent child in order to persuade her parent to reveal the whereabouts of a ticking bomb.

---

[38] Slote (2007) p. 7.
[39] Held (2006) p. 68.
[40] See Chapter One (§4.ii) for a more detailed discussion of culturally evolved heuristics.

The point here is not that these are *fatal* objections to utilitarianism: indeed, they are not. It may well be that the utilitarian can avoid such objections by offering a rule-based variant of their theory, for example. The point here is simply that these are *clearly* objections to utilitarianism. Yet this should not be at all obvious, at least from a strictly utilitarian perspective. If utilitarianism recommends certain actions as right in specific circumstances, and if one is a utilitarian, why should one not simply insist that those actions *really are* the right ones in those circumstances? Why should the fact that certain recommendations fail to accord with our intuitions tell against the theory which makes those recommendations, as opposed to telling against our intuitions themselves?

The same argument applies in equal measure to deontological theories of ethics. Kant's insistence on only acting upon maxims which were capable of being willed as a universal law of nature led him to the conclusion that one ought never, under any circumstances, to lie. He maintained that this conclusion held true even in the event of one's being asked by an axe-wielding maniac to disclose the whereabouts of her intended victim. Of course, Kantians may insist that this is not a fatal objection to their theory, and that Kant overlooked the fact that there is a difference between telling a lie and withholding information. Again, however, that is not the point. The point is that *were* Kant's theory in fact committed to such a claim, it would clearly count as a *problem* for that theory – *even by the lights of Kantian ethicists*.

Characterising expressed principles as the product of our operative intuitions allows us to understand this aspect of moral judgement. Not only should we *expect* our expressed principles, *qua* generalised rules-of-thumb, to fail to apply in certain situations; we can also understand why, *when* they so fail, we reject or modify our expressed principles instead of simply ignoring our operative intuitions. Expressed principles exist in order to promote and facilitate actions which conform to our operative principles. When they fail to do so, we recognise this on an intuitive level, and modify our course of action accordingly.

Such an interpretation accords with the care-ethical insistence on the appropriateness of contextualism, and places caring intentionality, albeit in the form of operative intuitions, at the heart of moral judgment. It is therefore possible to accommodate the existence of action-guiding expressed principles within a care-ethical framework.

Is there any reason to think that the interpretation I have just outlined is more than mere armchair speculation? In fact there is. Research into the importance of cognitive heuristics in moral decision making supports my interpretation. Cass Sunstein has argued that:

> [m]uch of everyday morality consists of simple, highly intuitive rules that generally make sense, but that fail in certain cases. [...] The problem comes when the generalisations are wrenched out of context and treated as freestanding or universal principles.[41]

Sunstein terms these intuitive rules "moral heuristics". As he defines them:

[41] Sunstein (2005) p. 531.

> heuristics are mental shortcuts used when people are interested in assessing a "target attribute" and when they substitute a "heuristic attribute" of the object, which is easier to handle.[42]

Sunstein remains agnostic with regard to the target attribute to which such heuristics relate. He notes that "it is possible to contend that a moral heuristic is at work without accepting any especially controversial normative claims".[43] However, for the present argument in favour of a care-ethical interpretation of moral discourse, the target attribute should be understood in terms of caring intentionality. The heuristic attributes, on the other hand, can be understood in a far more flexible way. They may be understood, for example, as acting in conformity with a specific duty, promoting certain good consequences, or being obedient to God's law. The specific heuristic attribute in question will perform its function as long as it both reliably co-occurs with caring motivations, and is in most circumstances easier to identify than those motivations. As long as one has a ready-to-hand list of duties, good consequences, or Divine commands, this will be the case for each of the three examples just provided.

Gigerenzer et al (2008) also endorse a heuristic interpretation of human decision making, though their focus is on non-moral heuristics. They argue that such heuristics can be acquired in a variety of ways. These include direct natural selection, social learning (through teaching or imitation) and individual learning (i.e. personal experience).[44] Gigerenzer et al emphasise that these cognitive heuristics enable an organism to respond quickly to complex environmental stimuli. More specifically, they do this in conditions in which a consciously reasoned response would be highly cognitively demanding or time consuming. Thus they are operations "that a mind can actually carry out under limited time and knowledge",[45] and which "perform intelligent guesses about unknown features of the world, based on uncertain indicators".[46]

Such research lends substantial support to my argument that expressed principles can be seen as subordinate to (i.e. as heuristic attributes of) operative principles. Yet even if this much is granted, the argument so far only gets the care-ethicist half-way to her desired conclusion. Why should operative principles be understood in specifically care-ethical terms? I have argued that operative principles share the care-ethicists' focus on the intentionality of agency. Still, this is not sufficient to allow us to favour care-ethics over other agent-centred theories of ethics, such as virtue ethics. In order to make the case for a specifically care-ethical interpretation of our operative ethical intuitions, I now turn to recent research into the evolution of social intelligence and theory of mind.

5. Social Intelligence and the Evolution of Empathy

[42] Sunstein (2005) p. 532.
[43] Sunstein (2005) p. 534.
[44] Gigerenzer et al (2008) p. 232.
[45] Gigerenzer and Goldstein (1996) p. 652.
[46] Gigerenzer and Goldstein (1996) p. 652.

The importance of empathy in recent accounts of caring has already been noted (in §3). Yet empathy is also a central aspect of research into the evolution of social intelligence. That human intelligence should have evolved to the high degree which it has is something of an evolutionary puzzle. It is tempting to suppose that intelligence is straightforwardly adaptive, such that the more you have, the better off you are likely to be. But this supposition becomes problematic when we consider that non-human animals, including other primates, are able to survive and reproduce with nothing like the level of intelligence displayed by humans. Human intelligence is therefore clearly not *necessary* for survival and replication, yet it is necessity which drives evolutionary change.

Nor does evolving additional intelligence come without costs. To begin with, brains require a lot of energy to develop and maintain: "[t]he brain's running costs are about 8 to 10 times as high, per unit mass, as those of skeletal muscle".[47] Furthermore, the increase in brain size which makes human levels of intelligence possible also generated unique adaptive problems. Occurring well after the transition to bipedal locomotion, the increasing size of the infant brain necessitated a corresponding increase in the width of the birth canal. However, there are limitations on the extent to which such widening is possible without severely compromising mobility. These limits were offset by the increasingly early birth of human babies. All humans are thus effectively born prematurely, "right on the limit at which it is possible for them to survive outside the womb".[48] This means that human infants are much more altricial at birth than are other primate infants, requiring much more care in order to have even a chance of survival. All of this, then, in order for our hunter-gather ancestors "to cope with what are essentially the same foraging decisions [as those made by squirrels]".[49] One would therefore expect significant selection pressure *against* evolving larger brains and the intelligence that goes with them. Why, then, did the hominini become so intelligent?

The answer to this puzzle lies in the unique selection pressures created by the social conditions in which human ancestors lived. Richard Alexander has argued that the ability to overcome a variety of social challenges would have been extremely selectively advantageous amongst ancestral humans.

> Individual reproductive success would depend increasingly on making the right decisions in complex social situations involving self, relatives, friends, and enemies. Critical choices would be aided by experience, and an intimate knowledge of the particular social environment. [...] In other words, because of the existence of groups intensely competitive with one another, and because of the complexity of social competition within groups evolving to be effective in intergroup competition, the human species in some sense became its own most important selective environment [...].[50]

Building on Alexander's thesis, Flinn et al (2005) provide additional details. They argue that the social intelligence hypothesis posited by Alexander cannot, if taken in isolation, explain the dramatic increase in human intelligence. This is because hominin social group sizes were not significantly different from those of other primates. Thus, they ask, "[w]hy were coalitions more

[47] Dunbar and Shultz (2007) p. 1344.
[48] Barrett et al (2002) p. 141.
[49] Dunbar and Shultz (2007) p. 1344.
[50] Alexander (1979) p. 214.

important, and more cognitively taxing, for our hominin ancestors than for any other species in the history of life?"[51]

In answering this question, Flinn et al draw on what they term the "ecological dominance-social competition" model. This account suggests that social competition will become important enough to act as the principal mechanism of evolutionary change only when a species is ecologically dominant. This is a state in which vulnerability to predation is reduced to a minimum, and limits on population size imposed by climatological factors or resource scarcity have been largely overcome. Once these environmental constraints on fitness become less significant, "the relative importance of selection from interactions with conspecifics"[52] increases dramatically. Once this change in relative selective importance takes place, a species can be said to be ecologically dominant.

The concept of ecological dominance allows Flinn et al to explain why the size of non-human primate social groups did not create the same selection pressure for hypertrophied intelligence, despite being comparable to the size of human social groups. Early humans' ability to develop increasingly sophisticated tools, and a behavioural adaptability which allowed them to draw on "variable and flexible subsistence strategies",[53] meant that they were able to become much more ecologically dominant than other primates.

As a result of this ecological dominance:

> Increasing linguistic and sociocognitive capacities were favoured because such skills allowed individuals to better anticipate and influence social interactions with other increasingly intelligent humans. [...] [T]he hominin social environment became an autocatalytic process, ratcheting up the selective advantage associated with the ability to anticipate the social strategies of other hominins and to mentally simulate and evaluate potential counterstrategies.[54]

The result of this autocatalytic process was the evolution of "uniquely sophisticated social problem-solving abilities, including TOM [i.e. theory of mind], language, consciousness, romantic love, and empathy".[55]

Why do we find empathy on this list? A sufficiently sophisticated theory of mind would enable its possessor to understand not only the intentions of one of her conspecifics, but also the limitations in that conspecific's knowledge, and how these limitations might be exploited. What does empathy do for us that a sophisticated theory of mind alone does not? The answer to this question is that it generates greater social cohesion. As Rizzolatti and Fogassi note, "if we see a person expressing pain, this does not mean we are forced to feel compassion [...] for them".[56] The evolution of a capacity for empathy, then, creates the possibility for an organism to do more than simply predict the behaviour of one of its conspecifics. Empathy:

---

[51] Flinn et al (2005) p. 13.
[52] Flinn et al (2005) p. 14.
[53] Flinn et al (2005) p. 15.
[54] Flinn et al (2005) p. 16.
[55] Flinn et al (2005) p. 37.
[56] Rizzolatti and Fogassi (2009) p. 193.

> allows an empathic reaction to another's emotional state. [...] It is this [...] that gives it the adaptive benefit of ensuring that organisms feel a drive to help each other.[57]

Without this additional emotional engagement, as Frans de Waal observes, "perspective-taking would be a cold phenomenon that could just as easily lead to torture as to helping".[58] De Waal therefore argues that empathy is "essential for the regulation of social interactions, coordinated activity, and cooperation toward shared goals".[59] For de Waal, the sort of empathy we typically associate with humans can be described as empathic perspective-taking. This is "a cognitive affair dependent on imagination and mental state attribution [...] in combination with emotional engagement".[60] The notion of empathic perspective-taking therefore suggests that, whilst conceptually distinct, empathy and theory of mind may often operate in conjunction with one another.

With the development of both a sophisticated theory of mind and of the capacity for empathy, then, the social stage is set for the emergence of something recognisable as genuine morality. Individuals with the advanced levels of social intelligence just discussed would have been able to make increasingly accurate internal representations of the mental states of their conspecifics. This would allow for large-scale in-group cooperation, as well as smaller-scale mutually advantageous interpersonal transactions.[61] Furthermore, such cooperative behaviours would have been promoted by an increasing tendency towards an empathetic engagement with the emotional states of conspecifics.

Of course, none of the foregoing is intended to imply the (frankly fantastical) conclusion that we cannot help but be nice toward one another. As de Waal notes, "[o]ne way to cognitively control empathy is to inhibit"[62] our tendency to identify with conspecifics. One way in which this might be done is by identifying, or even fabricating, some feature which can be used to class a conspecific as a non-group member. Human societies are remarkably adept at this task, as history shows. Nationality, skin colour, sex, sexual orientation, social class, and religious belief are only the most immediately striking examples of features that have been used to demarcate social groups, thereby facilitating empathy-blocking strategies of non-identification. Philip Kitcher suggests that the limitation on our tendency towards empathetic response is itself the result of natural selection. There are likely to be circumstances in which large gains in reproductive fitness will follow from non-altruistic actions. In such circumstances, it would pay an agent to disregard even the most long-standing alliances with conspecifics in order to secure those gains.

> Some features of recurring situations trigger an altruistic response at a particular intensity; in response to different features the animal is disposed to react with a different intensity of altruism, or perhaps to react with zero intensity.[63]

---

[57] Baron-Cohen (2009) pp. 215 – 216.
[58] de Waal (2008) p. 287.
[59] de Waal (2008) p. 282.
[60] de Waal (2008) p. 285.
[61] See the discussion of Joyce's "helping behaviours" in Chapter 3.
[62] de Waal (2008) p. 291.
[63] Kitcher (2011) p. 72.

Qualifications regarding the limitations on our tendency towards empathetic response aside, it is now possible to see that empathic perspective-taking, in conjunction with an increasingly sophisticated theory of mind, evolved as a means of facilitating cooperative endeavours at both an inter-personal and societal level.

It might be objected at this point that the sort of empathy which I have so far been discussing is of no use to the care-ethicist. As explained in §3, care-ethicists such as Slote and Noddings draw primarily on the notion of mediated associative empathy, whereas I have here been dealing exclusively with projective empathy. My response to this objection is to claim that these forms of empathy are likely to evolve hand-in-hand. If, as the empirical literature suggests, socially generated selection pressures were responsible for the emergence of empathy, then it is likely that both projective and mediated associative forms will have been simultaneously selected for. Consider the way in which a capacity for empathetic response is thought to have contributed to fitness. If empathy was selected in order to promote social cohesion, to allow bands of hunter-gatherers to work together more effectively, yet *also* to allow conspecifics to outwit other tribes in the competition for resources, then projective empathy alone would not have been up to the task. Effective competition and co-operation both require more than the ability to ask myself what I would do in your shoes. They also require the ability to predict what you will do *given your preferences and desires*. The former ability comes with projective empathy, whereas the latter falls in the remit of mediated associative empathy. The same selection pressures which produced the former are therefore highly likely to produce the latter as well.

The evolutionary narrative which I have sketched therefore sits comfortably with Noddings' description of caring as a mental state "directed toward the welfare, protection, or enhancement of the cared-for".[64] Given this, and given also the conclusion arrived at in the previous section, we can now see that all moral judgement can be conceived of in care-ethical terms.

My argument so far has been in favour of a metaethics of care: that an evolutionary form of care-ethics best articulates what it is that we are doing when we make moral judgements. Consequently, it is possible to object that even if this is the correct metaethical account of morality, we nevertheless should not embrace care-ethics as a means of conducting normative theory. For the purpose of this chapter, which is to set out a realist alternative to Joyce's evolutionary scepticism, this objection could be left unanswered. It seems somewhat disingenuous, however, to spend so long arguing in favour of a care-ethical account of morality, only to refuse to take a stance on the question of whether such an ethic could work in practice. In the following section then, I nail my colours to the wall and briefly set out an argument in favour of normative care-ethics.

6. Why We Should Endorse Normative Care-Ethics

An important objection to endorsing care-ethics (at least as described in this chapter) as a normative theory can be found in the literature on cognitive heuristics. It is a prominent feature of the accounts

[64] Noddings (2003) p. 23.

in Gigerenzer and Goldstein (1996) and Gigerenzer et al (2008) that cognitive heuristics are "ecologically rational".[65]

Ecological rationality is to be understood in contrast to the classical models of rationality which were popular during the Enlightenment. The classical model takes "standard statistical tools to be the normative and descriptive models of inference and decision making".[66] Gigerenzer and Goldstein (1996) argue that early research into cognitive heuristics preserved the normative standards of classical rationality, depicting the use of heuristics as a failure to attain the desired standard of rational thought. "Both views accept the laws of probability and statistics as normative, but they but disagree about whether humans can stand up to these norms".[67]

In claiming that cognitive heuristics typically exhibit ecological rationality, Gigerenzer and his co-authors mean to suggest that the laws of probability and statistics should not be considered normative models of human inference and decision making. Instead, they propose that human thought processes be conceived in terms of "bounded rationality".[68] That is, "the minds of living systems should be understood relative to the environment in which they evolved, rather than to the tenets of classical rationality".[69] The notion of bounded rationality draws attention to the pragmatic aspects of decision making. As previously noted, cognitive heuristics enable an organism to make decisions rapidly, and without access to all relevant information. The appropriate method of evaluating the decisions which heuristics prompt, however, is not to consider the nature of the deliberative processes involved. Rather, it is to assess whether the resulting action was in fact the means to the agent's desired end. Heuristics are designed to produce satisficing outcomes to deliberation, and are rational (more specifically, ecologically rational) precisely to the extent that they do so.

Sunstein also draws on the notion of ecological rationality in his discussion of moral heuristics. Accordingly, he acknowledges that:

> it is possible that in the domain of values as well as facts, real-world heuristics generally perform well in the real world – so that moral errors are reduced, not increased, by their use, at least compared to the most likely alternatives [...].[70]

Clearly, this possibility generates an objection to adopting care-ethics as a normative theory. According to this objection, if cultural evolution has led our expressed principles to become as widespread as they are, then we should assume this is for good reason: it is because they are ecologically rational. Engaging in contextualised care-ethical reasoning would be too cognitively demanding, and would leave us vulnerable to making potentially catastrophic moral misjudgements.

This is an important objection, but I do not believe that it is unanswerable. The first thing to point out in response is that the discussion in the previous sections has identified caring as the target attribute for which our expressed principles act as heuristic attributes. In the absence of some

---

[65] Gigerenzer et al (2008) p. 230.
[66] Gigerenzer and Goldstein (1996) p. 650.
[67] Gigerenzer and Goldstein (1996) p. 650.
[68] Gigerenzer and Goldstein (1996) p. 651.
[69] Gigerenzer and Goldstein (1996) p. 651.
[70] Sunstein (2005) p. 533.

such identification, attempting to ascertain the morally correct course of action would indeed be highly problematic. There would be no guidance beyond that which our expressed principles could provide, and these may sometimes conflict in philosophically familiar ways: promoting the best possible consequences may in some circumstances necessitate treating someone unjustly; doing the charitable thing may in some circumstances preclude doing the honest thing; upholding the rights of one individual may lead to a violation of the rights of another.

By identifying caring as the target attribute of moral judgement, we are better able to evaluate possible resolutions to the moral dilemmas which arise when our expressed principles come into conflict with one another. That is, care-ethical considerations give us a legitimate method to employ when trying to decide which of the conflicting principles provides unsound guidance in that instance (assuming, for the sake of argument, that we are not facing a tragic moral dilemma of the sort described by Rosalind Hursthouse[71]). This significantly reduces the risk that setting our expressed principles to one side will inevitably result in moral error.

Secondly, endorsing a normative ethics of care does not require us to eschew altogether the use of expressed principles. Saying this may seem to renege on the contextual approach to moral judgement endorsed by care-ethics, but this is not the case. As we have already seen, Noddings notes that some actions will very reliably prove to be incompatible with a caring attitude: "stealing is almost always wrong".[72]

We may make precisely the same judgement with regard to any number of our other expressed principles, be they consequential, deontological, or virtue ethical in content. Where the consequentialist, deontologist, or virtue ethicist will claim that their expressed principle constitutes the last word on moral judgement, however, the care-ethicist will deny this. Such principles should be thought of as *ceteris paribus* principles. Adopting a care-ethical perspective encourages us to be on the lookout for situations in which a dogmatic adherence to such principles threatens to lead us astray.

Care-ethics, then, is not vulnerable to the charge that it is too cognitively demanding to be of use in the real world. Adopting a care-ethical perspective allows us to diagnose the shortcomings of other normative theories, whilst accommodating the various insights which make such theories attractive to many philosophers. Certainly, we have a *prima facie* duty not to deceive someone, or not to steal from them: doing so is hardly ever compatible with a caring attitude towards that individual. Yet duty is not the whole of morality. Certainly, increasing the happiness of one's neighbour or loved one is a morally good thing: to do so is almost always to care for that person. Yet neither is utility the whole of morality. Each of these moral intuitions can be accommodated within an ethics of care, and an ethics of care offers the best prospect for normative guidance when these intuitions come into conflict with one another.

7. Evolutionary Care-Ethics and Moral Realism

[71] Hursthouse (2001) pp. 71 – 77.
[72] Noddings (2003) p. 93.

In this section, I argue that the evolutionary care-ethics which I have outlined above is compatible with a robust form of metaethical realism. Once this has been established, I will be able to provide a more detailed account, in §8, of how the evolutionary care-ethicist can respond to each of the aspects of Joyce's argument for a moral error theory. Before turning to that account, then, the present section shows that the care-ethical position which I have outlined vindicates characterising moral judgements as true or false. To that end, I start by discussing reference fixing accounts of morality, and how these have been used to support realist forms of sentimentalism. I argue that the account which I have sketched in this chapter ought to be understood as providing a reference-fixing account of moral terms. In §7.ii I argue that by giving a specifically evolutionary reference-fixing account, the theory which I have outlined is able to avoid moral relativism, and therefore does not face the same objections as the sentimentalist theories endorsed by Prinz (discussed in the previous chapter) and Slote.

(i) Sentimentalist Realism Revisited

Slote (2006) outlines a way of formulating a version of sentimental realism different from that found in Prinz. His account is worth noting as Slote places care-ethics firmly within the sentimentalist tradition:

> the idea of a morality of caring is [...] rather similar to Hutchesonian and Humean (normative) virtue ethics: benevolence and caring are both sentiments in the eighteenth-century sense, and an ethics of caring sees morality as based in feeling, or in motives that involve feeling, *rather than* in reason or rational principles.[73]

How, then, does Slote suggest a realist sentimentalism might be argued for?

According to his account, a realist sentimentalism should develop the "analogy between the sensory (e.g. colours) and the moral that projectivism and ideal-observer/response-dependent theories also rely on".[74] We have already discussed the difficulties which such approaches encounter with reference to Prinz's work in this vein. However, Slote argues that the problems traditionally associated with realist versions of projectivism can be avoided by developing the analogy between the sensory and the moral in terms of "reference fixing".[75]

Slote's remarks regarding this strategy are programmatic, but the guiding thought is that a (true) explanation of an occurrent sensation of red fixes the reference of the term "red," though without thereby specifying the semantic content of that term. Thus, "red" refers to all objects (and only those objects) with a specific reflectance profile, such that normal observers under normal conditions will experience objects with that reflectance profile as being red. That the reference of the term "red" can be fixed in this way does not, of course, specify the semantic content of the term

[73] Slote (2006) p. 225.
[74] Slote (2006) p. 236.
[75] Slote (2006) p. 236.

"red", as people who use this term do not typically intend to be understood as making claims about the reflectance profiles of particular objects. Slote argues that this approach:

> picks out a property in objects that is possibly identical with some surface feature of those objects (in relation to surrounding objects) and that is "rigidly" the same in all possible worlds (even those where different external properties normally cause us to have sensations of red).[76]

Of course, colour terms are vague, and the ascription of certain colour properties is notoriously contestable. This contestability creates serious problems for any attempt at fixing the reference of colour terms. This is not only because of the difficulty of identifying normal observers and specifying normal conditions, but also because two such observers may reasonably disagree about the colour of an object. These problems have been discussed by Joshua Gert, who argues that the vagueness of colour terms is not an objection to a reference-fixing strategy.

Gert's argument begins by positing a counterfactual possible world in which everyone has "precisely the same phenomenal color responses to the same objects, all the time".[77] In such a world, he argues, it would be reasonable to assume that colour terms have a straightforwardly referential function: they pick out objects in the visual field which have a specific sort of perceptual similarity to one another. This is in part because "uniform phenomenal response could be expected to figure in the process of language learning".[78] Uniformity of phenomenal response is not in itself sufficient to guarantee the emergence of specific colour terms, however. As a way buttressing his account, Gert therefore suggests that pragmatic reasons for identifying specific colours would also be a contributory factor in the emergence of colour terms. For example, there would be a pragmatic reason to develop a colour term for redness if all red berries were poisonous. Gert argues that in such a possible world:

> there is a fact of the matter, for any given object, whether it ought to be called by any given color word, so that the color words have, in a fairly straightforward sense, fixed extensions. Given our stipulated uniformity of response, such words will function almost like "marked with an X" or "marked with an O".[79]

Gert then goes on to argue that, by incrementally introducing other possible worlds which contain increasing vagueness in the extension of colour terms, it is possible to see that there nevertheless remains a fixed referent to those terms. In the second possible world, then, differences in light quality may cause observers to perceive the same colour differently, but two observers in the same conditions will have a uniform colour experience. Disagreements about the colour of certain objects will therefore occur, but these are quite straightforwardly resolvable as disagreement is produced solely by non-standard viewing conditions. Gert contends that colour terms continue to act as referring terms in this second possible world, in much the same way as terms relating to an object's shape.

---

[76] Slote (2006) p. 236.
[77] Gert (2007) p. 81.
[78] Gert (2007) p. 81.
[79] Gert (2007) p. 82.

> For while there is sometimes disagreement regarding the shape of an object, these disagreements can almost always be resolved by making sure that none of the observers are in non-standard circumstances for the observation of shape.

Gert iterates this strategy in subsequent possible worlds. Despite an increasing degree of vagueness in the applicability of colour terms, he argues, "[i]f they were referring terms in world 2, they continue to be referring terms in worlds 3a, 3b, 3c and so on".[80] In world 3c Gert introduces a significant degree of divergence in colour response, though he confines this divergence to a small percentage of the population. If the number of individuals who have divergent colour responses is small enough, Gert argues, "then it would be natural, if not inevitable, for the semantics of [colour terms] to remain more or less the same".[81] In such circumstances, Gert claims, we would expect to see the emergence of a term which identifies certain individual's responses as divergent from the statistical norm. This term would be expected to develop because "we do not want to rely on those [individuals] who regularly have"[82] divergent colour responses. Thus terms such as "colour-blind" enter the vocabulary. Once this occurs, Gert claims, the referential robustness of colour terms will be able to withstand a significant increase in the frequency of divergent colour responses.

> For there is no reason to think that the classification of responses as defective need be a strictly statistical matter. Rather, a host of pragmatic considerations will have their influence on the formation of the relevant concepts. [...] Thus, even if more and more people begin to suffer from [colour-blindness], the judgments of those who retain a fuller discriminatory capacity might plausibly continue to be regarded as authoritative.[83]

Gert argues that this same account can be used to fix the reference of terms such as "harmful" and "beneficial". Of course, with the latter terms there is difference in the nature of the response in question: colour terms elicit a phenomenal response, whereas terms such as "harmful" and "beneficial" elicit an affective response. Yet the argument which allows these to be identified as referring terms remains the same. Furthermore, the result of this argument, Gert holds, is a form of robust moral realism:

> just as agreement in phenomenal color responses facilitates the teaching of color words, agreement in affective responses facilitates the teaching of normative terms that then refer – directly – to a property of the things that elicit that response.[84]

I cannot provide a detailed critical response to Gert's theory here. For present purposes, it must suffice to use his account to flesh out some of the details in Slote (2006). Thus, when Slote claims that "statements about moral approval and disapproval fix the reference of moral properties/claims,"[85] we may understand him to be advocating the sort of approach defended by Gert.

---

[80] Gert (2007) p. 84.
[81] Gert (2007) p. 85.
[82] Gert (2007) p. 85.
[83] Gert (2007) p. 86.
[84] Gert (2007) p. 103.
[85] Slote (2006) p. 236. Parentheses removed.

> Moral claims would not, then, be *about* our reactive sentimental dispositions, but their (objective) reference would be fixed by facts about those dispositions, and such a view clearly deserves to be called a form of (sentimentalist) moral realism.[86]

Such an approach is commensurate with the care-ethical account of morality argued for in §§3 – 4 of this chapter. Phrasing that argument in the terms of the position outlined in this section, the claim being defended is that facts about the evolutionary aetiology of moral judgement can be used to fix the reference of moral discourse. This reference is to be identified with the caring intentionality of an agent's motive for action.

When construed in this way, moral judgements can be seen to be un-problematically truth-apt. When the claim is advanced that an agent acted immorally, there is a legitimate way to assess the putative truth of this claim. We will need to ascertain, for example, whether or not her actions were the result of empathetic perspective taking; whether she intended to promote the welfare of the patient in question; or whether any consequences of her action which were clearly incompatible with that welfare were foreseen by her. These considerations, it hardly needs saying, are intended to be illustrative rather than exhaustive.

In practice, of course, assessing the truth of any such description of an agent's intentional state may be an incredibly complex process, with little or no guarantee of success. Yet it should be uncontroversial that there is a judgement-independent, psychological fact-of-the-matter here, waiting, as it were, to be uncovered. That some such fact exists, however, is all that is needed to secure the claim that moral judgements are truth-apt.

(ii) Assessing the Realist Credentials of Reference-Fixing Approaches to Sentimentalism

As seen in the previous chapter, the realist pretentions of such an approach will likely only be found convincing if it can avoid the charge of implicit relativism. The sort of relativism which is a feature of Prinz's account is not implied by the version of care-ethics which I have outlined. The relativistic element of Prinz's theory was generated by his pluralistic account of basic moral values. The evolutionary credentials of the theory which I am defending allow me to replace Prinz's claim that there are diverse, irreconcilable, basic values with the claim that we all possess universally shared, care-ethical, operative intuitions. According to Prinz, it will be recalled, "we can figure out what our obligations are by figuring out what our moral beliefs commit us to".[87] If our moral beliefs commit each of us to the promotion of caring agency, then there is no threat from relativism.

Of course, it is possible that different cultures or social groups will, throughout the historical process of cultural evolution, have formed different and perhaps irreconcilable beliefs about *what constitutes* the promotion of caring agency. The view on offer here is therefore compatible with, and to an extent predicts, a certain amount of divergence between the moral norms of different cultures. It does not follow from this divergence that some of those norms *must* be morally wrong:

---

[86] Slote (2006) p. 237.
[87] Prinz (2009) p. 1.

perhaps they each promote caring relations in their own, unique way. But neither does it follow that any old set of expressed principles is as morally good as any other. Thus different sets of norms may be assessed in relation to how well they do *in fact* promote caring agency; under what conditions they do so (or fail to do so); and the extent to which they might be improved by incorporating elements from other systems of norms. Thus, *contra* Prinz, according to the theory on offer there need never be a point at which "rational debate is impossible".[88]

There is, however, another way in which the position I defend might be thought relativistic. This is found in the concern that it decides the truth of moral judgements on the arbitrary basis of what we happen to have evolved empathetically to relate to. Slote (2006) expresses this thought in the following passage:

> A more pressing issue for sentimentalism, however, arises from the way it makes moral distinctions follow distinctions in human empathy. [...] If, for example, white people tend to empathize more easily with other whites than they do with blacks, will this not permit and recommend morally repugnant forms of *discrimination*?[89]

This issue is most certainly a problem for the account which Slote provides. However, it is not a problem for the account which I defend in this chapter.

As seen earlier in relation to the issue of abortion (§2), Slote relies on differences in empathetic concern to ground what he sees as *prima facie* plausible moral intuitions. Thus, we do not have the same moral obligations towards both early- and late-stage foetuses, and this is simply because we cannot empathise with the former to the same extent as the latter. This reliance leads Slote seemingly to bite the bullet in the case of racism, claiming that (luckily for us) "studies [...] seem not to indicate any very marked or *basic* human tendency toward preference on the basis of (similar) skin color".[90] Just in case subsequent studies should overturn those findings, however, Slote goes on to note that some authors "have argued that a slight preference for those similar to one [...] may be morally acceptable and even desirable".[91]

Slote's remarks here show that something has gone badly wrong somewhere in his account. I contend that this error lies in a too exclusive focus on empathy, coupled with an insufficient emphasis on empathetic *caring*.

According to the view which I defend, it simply does not make *sense* to claim that "repugnant forms" of racial discrimination could ever be morally justified. This is a straightforward misapplication of the term "moral" and its cognates. Racial discrimination of the sort which worries Slote is quite obviously never caring: it simply is not motivated by a concern for the well-being or enhancement of the victim of discrimination. Even if it *were* the case that we had no empathetic concern for the well-being of those who were in some way physically different to us, it would not be possible to characterise this lack of concern as morally good. This is because morality is *fundamentally* a matter of *having* concerns for others.

---

[88] Prinz (2009) p. 125.
[89] Slote (2006) p. 237.
[90] Slote (2006) p. 237.
[91] Slote (2006) p. 258.

A different way of making the same point is to say that Slote's brand of care-ethics runs into trouble because it fixes the reference of the term "moral property" in the wrong way. Slote claims that moral properties are precisely those properties which are picked out by our occurrent sentimental dispositions (as does Prinz). He is therefore unable to claim that our sentimental dispositions might happen to be set up in such a way as to *fail* to pick out moral properties. Ergo, if people's sentiments happen to be slightly racist, then a little racism must be morally permissible. According to the evolutionary brand of care-ethics which I propose, the reference of the term "moral property" can instead be fixed by appealing to the Darwinian function of moral judgements. As I have argued, moral judgements evolved to track caring intentionality. That is their function. They perform that function just to the extent that they do in fact track caring intentionality. Thus, moral properties are those properties which pertain to an action's status as motivated-by-caring. Clearly, racist actions are not motivated by caring; as a result, such actions cannot be morally vindicated, even if our sentiments happen to be such that we are somewhat tolerant of racism.

The reference-fixing analysis which I propose has direct implications for how we approach normative ethics, as can be seen by a reconsideration of Slote's discussion of abortion. In dealing with this issue, it is important to note first that Slote's discussion does not seem to adequately characterise the moral relation which we typically take ourselves to stand in with regard to early-stage foetuses. For example, accepting Slote's account at face value makes the claim that prospective mothers care for their early-stage foetuses problematic. Clearly, however, such caring does take place. When a mother-to-be stops consuming alcohol, supplements her diet with folic acid and vitamin D, or quits smoking, we naturally explain this in terms of a provision of care for the early-stage foetus. Of course, there may be other ways of characterising this sort of caring relation, perhaps in terms of prospective care for the infant which the early-stage foetus may one day become. Yet such a move seems rather *ad hoc*, and even if it can be made to work, we have other moral intuitions which are not so easily explained away. Suppose there to be an individual who routinely uses early-stage abortion as a form of birth control. This agent, we may suppose, finds using contraception inconvenient. She does not empathetically engage with the early-stage foetus, and so does not have any moral qualms about her regular use of abortion clinics. It certainly seems that this agent's actions are morally objectionable. Yet if Slote's account is correct, it is hard to see why this should be the case.

When might we judge early-term abortion differently? Suppose a prospective mother is informed by her doctor that her child will be born severely disabled. The infant will live for a short time in agonising pain, shortly thereafter to die. It seems reasonable to say that in these circumstances, we would not judge early-stage abortion to be immoral (nor, perhaps, even late-stage abortion). This is most plausibly explained in terms of the compatibility of abortion in these circumstances with a caring attitude. The mother chooses to terminate the pregnancy *specifically because* the caring thing to do is to spare her child unnecessary, pointless suffering.

I take these scenarios to show that Slote's claim that we do not stand in a moral relation to early-stage foetuses is problematic. But what about the issue of late-stage abortion? Assuming, for the sake of argument, that this *is*, as Slote claims, intuitively less morally permissible than early stage abortion, how is this intuition to be explained? As seen in §3, Slote attempts to explain (and legitimate) this intuition by appealing to our capacity for empathy. We are able to empathise more with the late-stage foetus, and this generates in us a caring sentiment towards it. As I have argued, I

take this approach to be wrong-headed. A more plausible explanation of this intuition is to say that late-stage abortion is typically associated with a *withdrawal* of care. *Contra* Slote, we can, and do, care for early-stage foetuses, and by the time these become late-stage foetuses that care has been on-going for quite some time. Terminating a late-stage pregnancy can therefore seem to be a revocation of the care which has heretofore been on-going. We may therefore suspect that motives other than the child's welfare lie behind the decision to terminate the pregnancy. Of course, this need not in fact be the case. There are reasons for late-stage terminations which *can* reflect caring attitudes: certain severe brain abnormalities, for example, are only detectable after around twenty weeks of gestation. Our moral assessment of a particular instance of late-stage abortion might be made more positive if we learn that it was motivated by a desire to spare a child the suffering which such an abnormality would cause.

This account is of course speculative, but it does not rest on any particularly controversial assumptions. The point which I wish to emphasise, however, is that the evolutionary care-ethics which I propose has the ability to account for the moral intuitions which Slote identifies, but does so without relying on contingent facts about human empathetic responses. Consequently, it is not vulnerable to the counterintuitive normative implications of Slote's theory.

8. An Evolutionary Success Theory

In this final section, I show how an evolutionary perspective on care-ethics is able to rebut Joyce's argument for an error theory, and so therefore constitutes what Joyce terms an "evolutionary success theory".[92] I re-state the various aspects of Joyce's challenge to ethical naturalism, and explain how the metaethical perspective which I have outlined meets those challenges. I also show how my account is able to dovetail with much of what Joyce says about the nature of ordinary moral discourse. I take this to be a good thing, as I find much of what Joyce says about the commitments of ordinary moral discourse to be convincing.

(i) Moral Judgements are Truth-Apt

It will be recalled, firstly, that Joyce takes moral judgements to be assertoric. That is, they purport to make truth-apt claims about moral properties. I argued at length in §7.i that evolutionary care-ethics is truth-apt, and will simply re-state that claim here, without producing additional arguments in defence of it. I will, however, provide an example designed to make my claim to truth-aptness less abstract, and therefore (hopefully) more intuitively plausible.

Jill's belief that Jack acted immorally is truth-apt. She doubts that he is being honest when he tells her that he is unable to repay the money which she lent him. She knows that Jack thinks she has no real need of the money. She also knows that Jack has arranged to go out on a date later in

[92] Joyce (2001) p. 148.

the week, and will need some spending money. She takes Jack to have disregarded her interests in favour of his own, and to have insincerely promised to repay the loan on time. Jill takes Jack's actions towards her to be uncaring, and judges that he has acted badly on that account. But Jill is wrong. Unbeknownst to Jill, Jack had a sincere intention to repay his debt to her on time. He was prevented from doing so by Jill's friend Anna, who asked Jack to lend her some money to buy Jill a birthday present. Jack agreed to this, knowing that if Anna did not buy her a present, Jill would be crushed (Jill and Anna's friendship has been under a lot of strain recently, and both have invested a lot of time and effort in maintaining it). Jack believes that Jill's interests are much better served by his helping to maintain her friendship with Anna than by his repaying a small debt on time. Jack's reasons for not repaying Jill on time were not, then, what she took them to be. They were motivated by his concern for Jill's wellbeing, and if Jill were to become aware of this motivation, she would revise her moral judgement of Jack's actions. She would acknowledge that her previous judgement of Jack, being based on a misrepresentation of his psychology, was wrong. What made Jill's previous judgement false was its failure accurately to represent the moral properties which are pertinent to care-ethical moral deliberation; i.e. the facts about the intentional state which Jack was in when he acted.

(ii) The Dual Function of Moral Judgements

A second aspect of Joyce's account is his claim that moral judgements should be understood as having a dual function. That is, according to Joyce moral judgements "express both beliefs and conative non-belief states".[93] This view is compatible with the metaethical perspective which I am here defending. In Chapter Three I argued that moral judgements ought to be distinguished from moral beliefs, the latter being construed as exclusively cognitive. This distinction, when applied to evolutionary care-ethics, allows us to say the following.

It is possible for Jack to possess moral knowledge by which he is motivationally unmoved. For example, he may have the true belief that in circumstances C, the caring thing to do for Jill would be to ϕ. Jack may also truly believe that he is currently in C. Nevertheless, Jack need not thereby be motivated to ϕ. More specifically, he will not be so motivated if he does not care about Jill. It is important to note that we may still describe Jack's inaction in C as immoral. Facts about what constitutes a caring action fix the meaning of our moral terms. Jack's inaction is therefore morally culpable. Pointing this out to Jack will, in all likelihood, prompt him to ϕ. This is because Jack is a human agent, and it is overwhelmingly the case that human agents care about being moral. But if Jack is evil, and does not in fact care about being moral, then he will still not be motivated to ϕ.

Conversely, when Jack makes a moral *judgement* in favour of ϕ-ing in C, he is *ex hypothesi* motivated to ϕ in C. This is because moral judgements are conative as well as cognitive. In care-ethical terms, then, we can say that Jack not only has a true belief about ϕ's constituting a caring action in C, but that Jack also has a "commitment to behave in a fashion compatible with caring".[94]

---

[93] Joyce (2007/2006) p. 56.
[94] Noddings (2003) p. 91.

This commitment can be characterised as one of Bernard Williams' internal reasons, among which are included not only desires, but also "dispositions of evaluation [and] patterns of emotional reaction".[95] These dispositions of evaluation and patterns of emotional reaction are care-ethical in nature, and are generated by our evolved capacity to empathetically respond to the needs of our conspecifics.

(iii) Naturalising Moral Properties

Joyce claims that moral properties cannot be satisfactorily naturalised. His primary argument to that effect is that no natural property is capable of being categorically binding on an agent, irrespective of any of that agent's desires. So, against my proposal Joyce would object that even if moral discourse *has* evolved to track facts about agents' intentional states, no categorical moral obligations are thereby generated. This is true, as far as it goes. But Joyce looks in the wrong place for moral categoricity. As I argued in Chapter Three, categoricity is not a feature of moral properties *per se*; rather, it is generated by the conative aspect of moral *judgement*. To the extent that moral considerations matter more to us than non-moral considerations, moral judgements are felt to be categorically binding. Our moral reasons to ϕ motivationally "trump" any non-moral reasons that we may have not to ϕ. Thus, when we judge that we ought morally to ϕ, and then fail to ϕ (as clearly sometimes occurs), we act badly by our own lights.

(iv) Evolutionary Data and Epistemological Scepticism

The main argument which Joyce levels against moral realism is his epistemological claim that evolutionary data gives us good reason to doubt the existence of moral properties. Joyce gives an evolutionary aetiology of morality, which shows how natural selection could create creatures like us, who are inclined to make precisely the sort of moral judgements that we make. Joyce's aetiology makes no mention of moral properties or facts, however. This shows, he claims, that we need not appeal to the existence of moral properties when explaining our moral beliefs: it seems that we would believe as we do, whether or not moral properties actually exist. Accordingly, we lose any epistemic licence we thought we might have had for believing in the existence of moral properties.

Joyce's argument does what it sets out to do, but only as long as moral properties are conceived of in a particular way. If they are taken to exist totally independently of us, and so to be the sort of things which could antedate human existence, then Joyce's argument applies to them. He has given us good reason to doubt the existence of moral properties so conceived. But moral properties need not, and ought not, be thought of in such a way.

According to the evolutionary account which I have sketched in this chapter, moral properties *did* have a part to play in human evolution. *Qua* pro-social motivational states, the ability

[95] Williams (2002/1981) p. 105.

to detect moral properties was an important factor in the evolution of human intelligence. Much of the cognitive architecture with which our species is endowed was selected for on the basis of its contribution to fitness in an increasingly complex social environment. Once our ancestors became ecologically dominant, the ability to successfully navigate through that social environment came to be of paramount importance. Psychological traits such as theory of mind and a capacity for empathetic engagement were key products of this selection process. They allowed our ancestors to more accurately assess the intentions and reasons for action of their conspecifics. Actions which were believed to have been motivated by anti-social intentions were responded to with negative emotions, whereas actions which were believed to have been pro-socially motivated were responded to with positive emotions. These belief-emotion complexes constituted our ancestors, and our own, operative moral intuitions. Over time, cultural evolution produced more-or-less reliable moral heuristics, designed to track more easily the action-types which our operative intuitions respond to. These heuristics take the form of expressed moral principles, such as "do not lie", "always keep your promises", etc. At bottom, however, these heuristics serve the same evolutionary purpose as our operative moral intuitions, to which they are subordinate: they track intentional facts about agency.

The evolutionary account which I have offered in this chapter, then, gives us a reason to believe in the existence of moral properties. But why accept my account over Joyce's? The reason for doing so is simple: the evolutionary narrative which I endorse explains the emergence of morality in more detail than does Joyce's. The selection pressures which Joyce discusses (e.g. kin selection, reciprocal altruism etc.) are not unique to human beings: non-human animals are similarly subject to them. Yet non-human animals do not make moral judgements. These selection pressures are therefore clearly not sufficient for the evolution of morality. What more, then, is required for morality to emerge? It will be recalled that Joyce's answer to this question is "language". But if we now ask why *language* has evolved, and why human beings are the only species to have attained the level of cognitive sophistication required for the *development* of language,[96] Joyce will have to provide us with some additional answers. The most compelling answers with which he will be able to supply us will be those which draw on the ecological dominance hypothesis. That is, the more fully Joyce spells out his evolutionary aetiology, the more closely it will come to resemble the one which I have endorsed in this chapter. But that aetiology gives us good evolutionary reason to believe in moral properties. Joyce's claim that evolutionary data should undermine our epistemic credence in moral properties is therefore false.

(v) Summary

With each of Joyce's challenges to moral realism met, his argument for an error theory has no legs to stand on. There is much that Joyce gets right about the nature of morality, and its aetiology. Moral discourse is indeed committed to the statement of moral facts; our beliefs about morality are indeed influenced by our evolutionary development; and moral judgements are used to

[96] As opposed to *learning* language: this task has been achieved to various degrees by trained, non-human primates.

express our emotions, as well as our moral beliefs. Yet for all this, ordinary moral discourse need not be considered problematic. An evolutionary care-ethics vindicates moral discourse, and therefore constitutes an evolutionary success theory. According to such a theory, describing an action as moral or immoral is no more metaphysically problematic than saying "Mary ran out of the bathroom screaming because she was scared of that huge spider". We observe an action, and then infer the most likely psychological explanation of that action. Some of the psychological states whose existence we infer, we then identify either as moral or as immoral. These psychological states are not metaphysically special. They are moral if they constitute pro-social dispositions, immoral if they constitute anti-social ones. Making accurate judgements about these states is of course not always easy, but it is by no means a hopeless pursuit. Still less is it the case that this pursuit gives rise to a discourse which is fundamentally committed to metaphysical error.

9. Conclusion

This chapter has outlined a realist response to Joyce's evolutionary argument for an error theory. Moral judgements are best understood care-ethically. That is, when we make a moral judgement we are attempting to accurately assess the intentions which produced an agent's actions. Roughly speaking, if an agent's intentions were to promote the well-being of a conspecific, then that agent's intentions were morally good. Empirical data from moral psychology supports the claim that judgements about agents' intentional states ground our moral intuitions. Independent evolutionary data places the human capacities for theory of mind and empathetic-perspective-taking at the centre of our moral thought. Taken together, these different empirical sources support a specifically care-ethical interpretation of our moral intuitions.

Evolutionary care-ethics is both realist and non-relativistic. Moral judgements are straightforwardly truth-apt, being true if they accurately represent an agent's intentions, and false if they do not. Furthermore, by combining an evolutionary approach with a reference-fixing strategy, a realist care-ethic is not committed to the endorsement of contingent, occurrent sentiments. The account which I have defended constitutes a clear improvement on the theories of Prinz and Slote in this regard.

Joyce's analysis of moral discourse, and his evolutionary aetiology of morality, both contain important truths. The account which I have outlined in this chapter can accommodate those truths, but it is also able to make room for the existence of un-problematically naturalistic moral properties, and to explain how we have epistemic access to them. Joyce's evolutionary argument for an error theory does not, therefore, succeed.

Conclusion to the Thesis

This thesis has shown that a realist reply can be made in response to Richard Joyce's evolutionary argument for a moral error theory.

Chapter One set the scene by discussing modern evolutionary interpretations of human behaviour. I highlighted the central tenets of E. O. Wilson's sociobiological research program, and of those of its historical successor, evolutionary psychology. I showed how such an approach can inform research into cultural anthropology, and that evolutionary data is therefore a genuinely informative resource when attempting to understand human behaviour. I then considered some objections to evolutionary interpretations of human behaviour, before going on to endorse developmental systems theory as a way of meeting those objections.

Chapter Two showed how looking at evolutionary theories of ethics from a developmental perspective can help to avoid conceptual confusion over what constitutes an evolutionary theory of ethics. This distinction is typically made from a non-developmental standpoint, drawing on intuitions either about: (i) the extent to which morality is "in" our genes; or (ii) the likelihood that our moral beliefs were the target of natural selection. From a developmental perspective, these questions can respectively be shown to be: (i) conceptually confused and (ii) irrelevant. Emphasising the distinction between moral philosophy and moral psychology, I claimed that ethical theories are evolutionary only if they attempt to draw normative or metaethical conclusions directly from evolutionary data. Claiming that evolutionary processes influenced our moral beliefs is not, in itself, sufficient to make one an evolutionary ethicist, though it does make one an evolutionary moral psychologist.

In Chapter Three I introduced Joyce's argument for an error theory. Joyce's argument uses evolutionary data to undermine the epistemic warrant of (what he takes to be) our ordinary moral beliefs. According to Joyce's analogy with the battle of Waterloo, evolution's influence on our psychology acts as the equivalent of a belief pill, which we have all unknowingly been slipped: the pill determines that we believe there to be moral facts, and does so irrespective of whether such facts actually exist. Knowing, as we now do, that our psychologies have been thus influenced by evolution, we are no longer justified in continuing to believe in the existence of moral facts. Furthermore, Joyce argues, it is unlikely that moral facts can be accommodated in a naturalistic universe. This is because they must possess the property of being intrinsically motivating, and such a property is metaphysically queer. Given that ordinary moral discourse is committed to the existence of such facts, Joyce argues, it is committed to an error.

Having set out Joyce's argument, I considered some possible objections to it. I argued that each of these objections, for different reasons, fails. However, I took one of them, made by Jamie Dreier, as the inspiration for an argument against Joyce's claim that moral properties must be conceived of as intrinsically motivational. This allowed me to refute Joyce's argument that moral properties cannot be cashed out naturalistically. Having done so, it still remained to find a naturalistic account of moral properties which would be capable of motivating belief in them.

I began this search in Chapter Four, with a discussion of Philippa Foot's neo-Aristotelian virtue theory. According to Foot's theory, possessing virtues is akin to possessing good health, and to

describe a behavioural trait as morally good is to engage in the same type of evaluation that we use to describe someone's eyesight as good. That is, in both instances, the term "good" serves to highlight that the trait in question contributes to the flourishing of the organism that possesses it. If moral properties are conceived of as properties which are conducive to an agent's flourishing, then evolutionary data need not undermine our belief in them. An evolutionary narrative can explain how an organism's species came to flourish in the way in which it does.[1] But such an explanation need not undermine the coherence of the claim that the possession of certain traits, and not others, promote that flourishing. Foot's position therefore has *prima facie* potential as a reply to Joyce. Unfortunately for the moral realist, however, that potential is not realisable. Foot's theory faces insurmountable philosophical difficulties, these being generated by her commitment to *eudaimonism*, and by her welfarist conception of function. The Darwinian account of the virtues proposed by Jonathan Haidt and Craig Joseph is similarly unable to block Joyce's argument. Although their theory does not face the same difficulties as Foot's, it is explicitly a descriptive, non-normative theory. As such, it does not attempt to vindicate talk of moral properties, and so plays directly into the hands of Joyce's error theory.

In Chapter Five I discussed Jesse Prinz's realist sentimentalism. Prinz's metaethical views are compatible with the Darwinian moral psychology endorsed by Haidt and Joseph, being partly derived from that psychology. His account of moral properties therefore suggests a way of supplementing Haidt and Joseph's Darwinian virtue theory, such as to make it metaethically realist. According to Prinz, and in line with Haidt and Joseph's account, moral judgements are emotionally constituted. Moral properties can nevertheless be said to exist, Prinz argues, as long as these are understood as being those properties which are causally responsible for the elicitation of a moral emotion. Prinz's metaethics is deeply problematic, however. His account of basic values is inconsistent with his theory of emotion, according to which emotions, including the moral emotions, are non-cognitive embodied appraisals. Moreover, his metaethical realism is unable to accommodate all of the features of realist moral discourse.

In Chapter Six I argued that a more attractive alternative to Prinz's theory can be found in an evolutionary perspective on care-ethics. Such an approach can accommodate several of Prinz's insights into realist sentimentalism, including his notion of basic values (when these are construed care-ethically), and his claim that normative conclusions follow from making our basic values explicit; i.e. by more explicitly articulating "what our moral beliefs commit us to".[2] From an evolutionary care-ethical perspective, I argued that our moral intuitions are generated by operative care-ethical principles, which normatively "trump" the expressed moral principles to which we consciously subscribe. I drew support from this argument from evolutionary data regarding the evolution of social intelligence, showing that ancestral human's ecological dominance created selection pressures for theory of mind and empathetic perspective taking. These, I argued, underlie our capacity for making moral judgements, and allow us to provide a reference-fixing account of moral discourse. I then went on to show that evolutionary care-ethics is a form of metaethical realism, and that adopting such a perspective gives us good evolutionary reason to believe in the

[1] Though evolutionary explanations do not, according to Foot, determine what *constitutes* an organism's flourishing.
[2] Prinz (2007) p. 1.

existence of moral properties. Joyce's epistemological argument for an error theory was thereby refuted.

Chapter Six showed that a satisfactory realist reply to Joyce is available. The evolutionary brand of care-ethics which it presented was only very roughly sketched, however. The thesis therefore points to a number of interesting areas for further research. In the remainder of this conclusion, I will briefly gesture to three of these.

(i) The Relation of Caring to Virtue

Firstly, the relation between our operative principles, *qua* care-ethical intuitions, and our expressed moral principles needs further specification. With specific reference to the topics addressed in this thesis, such specification could fruitfully relate care-ethical intuitions to our conception of the virtues. It has been argued that caring should be characterised as one of the virtues, and consequently that care-ethics should be thought of as a part of virtue theory, and not as constituting an independent normative ethic.[3]

A different way of relating caring to the virtues emerges from the perspective which I advocate in this thesis. Rather than analysing care as one of the virtues, it becomes possible to use the notion of caring to provide a normative foundation for the virtues. According to such a conception, each of the virtues picks out a particular way in which a caring attitude can be psychologically manifested. Insofar as the virtues can be taken to constitute guides to action, they can also be said to constitute heuristic attributes of moral discourse, designed to track the target attribute of caring intentionality. That is to say, the virtues are normative because they are dispositions which promote caring.

This way of conceiving the virtues is attractive, as it allows them to fit into a strictly Darwinian conceptual framework. Thus, the problems associated with neo-Aristotelian accounts of the virtues (as discussed in relation to Foot, in Chapter 4) do not arise. A care-centric account of the virtues is therefore likely to prove more philosophically defensible than traditional, neo-Aristotelian approaches.

(ii) Evaluating the Reliability of Moral Intuitions

The reliability of our moral intuitions is a subject of lively debate. There is a growing amount of empirical research which suggests that our intuitions are susceptible to a large variety of influences, many of which seem irrelevant to normative judgement. These range from framing effects to the tidiness of one's desk.[4] However, the judgement that certain influences on our

---

[3] See, for example, Halwani (2003).
[4] See Copp (2012) for discussion.

intuitions are normatively irrelevant is itself a normative intuition, albeit one which we instinctively trust. But if our intuitions can be manipulated as easily as research suggests, how can we be warranted in trusting *any* of them, including our epistemic intuitions about which of our intuitions are trustworthy?

One of the ways in which to settle this question is to provide an aetiological account of each of our moral intuitions. Some of those intuitions might then be shown to be occasioned by morally relevant features of situations, with others being occasioned by morally irrelevant features. The reference-fixing account of moral terms which I advocate in Chapter Six makes it possible to distinguish morally relevant features of situations from morally irrelevant ones. Succinctly, it could be argued that only those intuitions which are elicited by care-ethical considerations are to count as normatively reliable.

Suppose we have an intuition that Jack has acted somewhat badly. This might be because of the intentional state which we believe Jack to have been in when he acted. If so, then given the reference-fixing account of moral discourse which I have provided, this intuition is normatively relevant. It should play a part in a more detailed consideration of the moral worth of Jack's action. Suppose, however, that the dirtiness of our immediate surroundings was responsible for the elicitation of our intuition about Jack. How we cash out "responsible" here of course stands in need of further specification, but for now we can suppose that it would draw on something like Prinz's analysis of moral disgust (discussed in Chapter Five). In this latter case, our intuition would *not* be normatively relevant. This is because an instinctive wariness of microbial contamination did not feature in the reference-fixing account of moral discourse which I endorsed.

Much more needs to be said about each of these issues. Nevertheless, an evolutionary reference-fixing strategy has the potential to non-arbitrarily distinguish between those moral intuitions by which we should be influenced, and those by which we should not.

(iii) The Normative Applications of Evolutionary Care-Ethics

By identifying caring agency as the target attribute of both deontological and consequentialist moral heuristics, it becomes possible to resolve conflict between deontological and consequentialist moral intuitions. This can be done by ascertaining which, if either, of these intuitions most effectively promotes or maintains caring relations, given the details of the situation in which each such conflict arises.

Such an approach is clearly applicable, at the level of applied ethics, to the current debate surrounding voluntary euthanasia, and the conditions in which assisted suicide ought to be permissible. It is also applicable to problems in normative ethics, in which it could be used to settle debate over the normative relevance of the distinction between intended and foreseen consequences, and how this distinction is to be applied to the normative discussion of issues such as abortion.

(iv) Closing Remarks

The directions for further research that I have gestured to show that the evolutionary approach to care-ethics outlined in this thesis has the potential to be developed in a number of interesting directions. This potential is additional to the further research necessary to articulate more fully the theory itself. The prospects for such a theory are therefore promising, creating as they do the potential not only to articulate a metaethically realist normative framework, but to derive that framework from the best available interpretations of how and why we evolved to be moral animals.

Bibliography

- Aldrich, Howard E.; Hodgson, Geoffrey M.; Hull, David L.; Knudsen, Thorbjørn; Mokyr, Joel; and Vanberg, Victor J. 2008. "In Defence of Generalized Darwinism," in *Journal of Evolutionary Economics*, 18, pp. 577 – 596.
- Alexander, Richard D. 1979. *Darwinism and Human Affairs*. London: Pitman Publishing Limited.
- Alexander, Richard D. 1987. *The Biology of Moral Systems*. Hawthorne: Aldine de Gruyter.
- Anscombe, G. E. M. 1981a. "On Promising and its Justice," in *Collected Philosophical Papers*, Vol. 3, pp. 10 – 21. Minneapolis: University of Minnesota Press.
- Anscombe, G. E. M. 1981b. "Rules, Rights and Promises," in *Collected Philosophical Papers*, Vol. 3, pp. 92 – 103. Minneapolis: University of Minnesota Press.
- Aristotle. 2002. *Nicomachean Ethics*, translated by Christopher Rowe. New York, NY: Oxford University Press.
- Ayala, Francisco. 2006. "Biology to Ethics: An Evolutionist's View of Human Nature," in Giovanni Boniolo and Gabriele de Anna (eds.) *Evolutionary Ethics and Contemporary Biology*, pp. 141 – 158. New York, NY: Cambridge University Press.
- Baillie, James. 2000. *Hume on Morality*. London: Routledge.
- Baron-Cohen, Simon. 2007. "The Evolution of Empathizing and Systemizing: Assortative Mating of Two Strong Systemizers and the Cause of Autism," in R. I. M. Dunbar and Louise Barrett (eds.) *The Oxford Handbook of Evolutionary Psychology*, pp. 213 – 226. New York, NY: Oxford University Press.
- Barrett, Louise; Dunbar, Robin; and Lycett, John. 2002. *Human Evolutionary Psychology*. New York, NY: Palgrave Macmillan.
- Blackburn, Simon. 1993. *Essays in Quasi-Realism*. Oxford: Oxford University Press.
- Blackburn, Simon. 1995. "The Flight to Reality," in Rosalind Hursthouse, Gavin Lawrence, and Warren Quinn (eds.) *Virtues and Reasons: Philippa Foot and Moral Theory*, pp. 35 – 56. New York, NY: Oxford University Press.
- Bostock, David. 2000. *Aristotle's Ethics*. New York, NY: Oxford University Press.
- Buss, David M.; Haselton, Martie G.; Shackelford, Todd K.; Bleske, April L.; and Wakefield, Jerome C. 1998. "Adaptations, Exaptations, and Spandrels," in *American Psychologist*, Vol. 53 No.5, pp. 533 – 548.
- Carroll, Joseph. 1995. "Evolution and Literary Theory," in *Human Nature*, Vol. 6 No.2, pp. 119 – 134.
- Christensen, David. 1994. "Conservatism in Epistemology," in *Nous*, Vol. 28 No. 1, pp. 69 – 89.
- Cohen, Lawrence E.; and Machalek, Richard. 1988. "A General Theory of Expropriative Crime: An Evolutionary Ecological Approach," in *American Journal of Sociology*, Vol. 94 No. 3, pp. 465 – 501.
- Copp, David. 2001. "Realist-Expressivism: A Neglected Option for Moral Realism," in *Social Philosophy and Policy*, 18, pp. 1 – 43.
- Copp, David. 2009. "Realist-Expressivism and Conventional Implicature," in Russ Shafer-Landau (ed.) *Oxford Studies in Metaethics* Vol. 4, pp. 167 – 202. New York, NY: Oxford University Press.

- Copp, David. 2012. "Experiments, Intuitions, and Methodology in Moral and Political Theory," in Russ Shafer-Landau (ed.) *Oxford Studies in Metaethics* Vol. 7, pp. 1 – 36. Oxford: Oxford University Press.
- Cosmides, Leda; and Tooby, John. 1992. "Cognitive Adaptations for Social Exchange," in Jerome H. Barkow, Leda Cosmides, and John Tooby (eds.) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, pp. 163 – 228. New York, NY: Oxford University Press.
- Cosmides, Leda; Tooby, John; and Barkow, Jerome H. 1992. "Introduction: Evolutionary Psychology and Conceptual Integration," in Jerome H. Barkow, Leda Cosmides, and John Tooby (eds.) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, pp. 3 – 15. New York, NY: Oxford University Press.
- Cuneo, Terence. 2006. "Saying what we Mean: An Argument against Expressivism," in Russ Shafer-Landau (ed.) *Oxford Studies in Metaethics* Vol. 1, pp. 35 – 72. New York, NY: Oxford University Press.
- D'Arms, Justin; and Jacobsen, Daniel. 2000. "Sentiment and Value," in *Ethics*, Vol. 110 No.4, pp. 722 – 748.
- D'Arms, Justin; and Jacobsen, Daniel. 2006. "Sensibility Theory and Projection," in David Copp (ed.) *The Oxford Handbook of Ethical Theory*, pp. 186 – 218. New York, NY: Oxford University Press.
- Dawkins, Richard. 1976. *The Selfish Gene*. Oxford: Oxford University Press.
- Dawkins, Richard. 1982. *The Extended Phenotype: The Long Reach of the Gene*. Oxford: Oxford University Press.
- de Waal, Frans. 2006. *Primates and Philosophers: How Morality Evolved*. Princeton: Princeton University Press.
- de Waal, Frans. 2008. "Putting the Altruism Back into Altruism: The Evolution of Empathy," in *Annual Review of Psychology*, 59, pp. 279 – 300.
- Doyle, James. 2000. "Moral Rationalism and Moral Commitment," in *Philosophy and Phenomenological Research*, Vol. 60 No. 1, pp. 1 – 22.
- Dreier, Jamie. 2010. "Mackie's Realism: Queer Pigs and the Web of Belief," in Richard Joyce and Simon Kirchin (eds.) *A World Without Values*, pp. 71 – 86. Dordrecht: Springer.
- Dretske, Fred. 1986. "Misrepresentation," in Radu J. Bogdan (ed.) *Belief: Form, Content and Function*, pp. 17 – 36. Oxford: Oxford University Press.
- Dunbar, R. I. M; and Schultz, Suzanne. 2007. "Evolution in the Social Brain," in *Science*, 317, pp. 1344 – 1347.
- Durham, William H. 1991. *Coevolution: Genes, Culture, and Human Diversity*. Stanford, CA: Stanford University Press.
- Foot, Philippa. 2001. *Natural Goodness*. New York, NY: Oxford University Press.
- Foot, Philippa. 2002a. "Morality as a System of Hypothetical Imperatives," in *Virtues and Vices and Other Essays in Moral Philosophy*, pp. 157 – 174. New York, NY: Oxford University Press.
- Foot, Philippa. 2002b. "Introduction," in *Moral Dilemmas and Other Topics in Moral Philosophy*, pp. 1 – 3. New York, NY: Oxford University Press.
- Foot, Philippa. 2002c. "Preface to the 2002 Edition," in *Virtues and Vices and Other Essays in Moral Philosophy*, pp. ix – x. New York, NY: Oxford University Press.

- Foot, Philippa. 2002d. "Rationality and Virtue," in *Moral Dilemmas and Other Topics in Moral Philosophy*, pp. 159 – 174. New York, NY: Oxford University Press.
- Flinn, Mark V.; Geary, David C.; and Ward, Carol V. 2005. "Ecological Dominance, Social Competition, and Coalitionary Arms Races: Why Humans Evolved Extraordinary Intelligence," in *Evolution and Human Behavior*, 26, pp. 10 – 46.
- Gert, Joshua. 2007. "Cognitivism, Expressivism, and Agreement in Response," in Russ Shafer-Landau (ed.) *Oxford Studies in Metaethics*, Vol. 2, pp. 77 – 110. New York, NY: Oxford University Press.
- Gibbard, Allan. 1990. *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Cambridge, MA: Harvard University Press.
- Gibbard, Allan. 2003. *Thinking How to Live*. Cambridge, MA: Harvard University Press.
- Gigerenzer, Gerd; and Goldstein, Daniel G. 1996. "Reasoning the Fast and Frugal Way: Models of Bounded Rationality," in *Psychological Review*, Vol. 103 No. 4, pp. 650 – 669.
- Gigerenzer, Gerd; Goldstein, Daniel G.; and Hoffrage, Ulrich. 2008. "Fast and Frugal Heuristics are Plausible Models of Cognition: Reply to Dougherty, Franco-Watkins, and Thomas (2008)," in *Psychological Review*, Vol. 115 No. 1, pp. 230 – 239.
- Gilligan, Carol. 1982. *In a Different Voice: Psychological Theory and Women's Development*. Cambridge, MA: Harvard University Press.
- Gould, Stephen J. 1980. "Sociobiology and Human Nature: A Postpanglossian Vision," in Ashley Montagu (ed.) *Sociobiology Examined*, pp. 283 – 290. New York, NY: Oxford University Press.
- Gould, Stephen J. 1991. "Exaptation: A Crucial Tool for an Evolutionary Psychology," in *Journal of Social Issues*, Vol. 47 No. 3, pp. 43 – 65.
- Gould, Stephen J.; and Lewontin, Richard C. 1979. "The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme," in *Proceedings of the Royal Society of London,* B, 205, pp. 581 – 598.
- Gould, Stephen J.; and Vrba, Elizabeth S. 1982. "Exaptation – A Missing Term in the Science of Form," in *Paleobiology*, Vol. 8 No. 1, pp. 4 – 15.
- Griffiths, Paul E; and Gray, Russell D. 1994. "Developmental Systems and Evolutionary Explanation," in *The Journal of Philosophy*, Vol. 91 No. 6, pp. 277 – 304.
- Hacker-Wright, John. 2009. "What is Natural About Foot's Ethical Naturalism?" in *Ratio*, Vol. 22, pp. 308 – 321.
- Haidt, Jonathan. 2001. "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment," in *Psychological Review*, Vol. 108 No. 4, pp. 814 – 834.
- Haidt, Jonathan. 2004. "The Emotional Dog Gets Mistaken for a Possum," in *Review of General Psychology*, Vol. 8 No. 4, pp. 283 – 290.
- Haidt, Jonathan; and Joseph, Craig. 2004. "Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues," in *Daedalus*, Vol. 133 No. 4, pp. 55 – 66.
- Halwani, Raja. 2003. "Care Ethics and Virtue Ethics," in *Hypatia*, Vol. 18 No. 3, pp. 161 – 192.
- Hamilton, William D. 1964. "The Genetical Evolution of Social Behaviour. II," in *Journal of Theoretical Biology*, 7, pp. 17 – 52.
- Hauser, Marc D. 2006. *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*. London: Abacus.

- Hauser, Marc D.; Cushman, Fiery; Young, Liane; Jin, R. Kang-Xing; and Mikhail, John. 2007. "A Dissociation Between Moral Judgments and Justifications," in *Mind & Language*, Vol. 22 No. 1, pp. 1 – 21.
- Held, Virginia. 2006. *The Ethics of Care: Personal, Political, and Global*. New York, NY: Oxford University Press.
- Hume, David. 1978. *A Treatise of Human Nature*. New York, NY: Oxford University Press.
- Hursthouse, Rosalind. 1999. *On Virtue Ethics*. New York, NY: Oxford University Press.
- Jablonka, Eva; and Lamb, Marion J. 2005. *Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*. Cambridge, MA: The MIT Press.
- Joyce, Richard. 2001. *The Myth of Morality*. Cambridge: Cambridge University Press.
- Joyce, Richard. 2006. *The Evolution of Morality*. Cambridge, MA: The MIT Press.
- Joyce, Richard. 2010. "Patterns of Objectification," in Richard Joyce and Simon Kirchin (eds.) *A World Without Values*, pp. 35 – 54. Dordrecht: Springer.
- Joyce, Richard. 2011. "The Accidental Error Theorist," in Russ Shafer-Landau (ed.) *Oxford Studies in Metaethics* Vol. 6, pp. 153 – 180. New York, NY: Oxford University Press.
- Kahane, Guy. 2011. "Evolutionary Debunking Arguments," in *Nous*, Vol. 45 No. 1, pp. 103 – 125.
- Keller, Evelyn Fox. 2010. *The Mirage of a Space between Nature and Nurture*. Durham, NC/London: Duke University Press.
- Kitcher, Philip. 2011. *The Ethical Project*. Cambridge, MA: Harvard University Press.
- Konner, Melvin. 2010. *The Evolution of Childhood: Relationships, Emotion, Mind.* Cambridge, MA: The Belknap Press of Harvard University Press.
- Laland, Kevin N.; and Brown, Gillian R. 2006. "Niche Construction, Human Behavior, and the Adaptive-Lag Hypothesis," in *Evolutionary Anthropology*, Vol. 15 pp. 95 – 104.
- Lazarus, Richard S. 1991. *Emotion and Adaptation*. New York, NY: Oxford University Press.
- Lewens, Tim. 2010. "Foot Note," in *Analysis*, Vol. 70 No. 3, pp. 468 – 473.
- Lieberman, Debra. 2008. "Moral Sentiments Relating to Incest: Discerning Adaptations from By-products," in Walter Sinnott-Armstrong (ed.) *Moral Psychology Vol. 1. The Evolution of Morality: Adaptations and Innateness*, pp. 165 – 190. Cambridge, MA: The MIT Press.
- Lumsden, Charles J.; and Wilson, Edward O. 1981. *Genes, Mind, and Culture*. Cambridge, MA: Harvard University Press.
- Mackie, John L. 1977. *Ethics: Inventing Right and Wrong*. London: Penguin.
- Mackie, John L. 1980. *Hume's Moral Theory*. New York, NY: Routledge.
- Mayr, Ernst. 2001. *What Evolution Is*. London: Phoenix.
- Millum, Joseph. 2006. "Natural Goodness and Natural Evil," in *Ratio*, Vol. 19, pp. 199 – 213.
- Milton, John. 2000. *Paradise Lost*. New York, NY: Penguin.
- Montagu, Ashley. 1980. "Introduction," in Ashley Montagu (ed.) *Sociobiology Examined*, pp. 3 – 14. New York, NY: Oxford University Press.
- Morton, Luise H.; and Foster, Thomas R. 1991. "Goodman, Forgery, and the Aesthetic," in *The Journal of Aesthetics and Art Criticism*, Vol. 49 No. 2, pp. 155 – 159.
- Noddings, Nel. 1984. *Caring: A Feminine Approach to Ethics and Moral Education*. Berkeley, CA: University of California Press.
- Okasha, Samir. 2006. *Evolution and the Levels of Selection*. New York, NY: Oxford University Press.

- Oyama, Susan. 1985. *The Ontogeny of Information: Developmental Systems and Evolution*. Durham, NC: Duke University Press.
- Oyama, Susan. 2000a. "Causal Democracy and Causal Contributions in Developmental Systems Theory," in *Philosophy of Science*, Vol. 67, Supplement. Proceedings of the 1998 Biennial Meetings of the Philosophy of Science Association. Part II: Symposia Papers, pp. S332 – S347.
- Oyama, Susan. 2000b. *Evolution's Eye: A Systems View of the Biology-Culture Divide*. Durham, NC: Duke University Press.
- Oyama, Susan. 2001. "Terms in Tension: What Do You Do When All the Good Words Are Taken?" in Susan Oyama, Paul E. Griffiths, and Russell D. Gray (eds.) *Cycles of Contingency: Developmental Systems and Evolution*, pp. 177 – 194. Cambridge, MA: The MIT Press.
- Phillips, David. 2010. "Mackie on Practical Reason," in Richard Joyce and Simon Kirchin (eds.) *A World Without Values*, pp. 87 – 100. Dordrecht: Springer.
- Pinker, Steven. 2002. *The Blank Slate: The Modern Denial of Human Nature*. London: Penguin.
- Pinker, Steven; and Bloom, Paul. 1992. "Natural Language and Natural Selection," in Jerome H. Barkow, Leda Cosmides, and John Tooby (eds.) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, pp. 451 – 494. New York, NY: Oxford University Press.
- Prinz, Jesse J. 2002. *Furnishing the Mind: Concepts and their Perceptual Basis*. Cambridge, MA: The MIT Press.
- Prinz, Jesse J. 2004. *Gut Reactions: A Perceptual Theory of Emotion*. New York, NY: Oxford University Press.
- Prinz, Jesse J. 2005. "Are Emotions Feelings?" in *Journal of Consciousness Studies*, Vol. 12 No. 8 – 10, pp. 9 – 25.
- Prinz, Jesse J. 2006. "The Emotional Basis of Moral Judgments," in *Philosophical Explorations*, Vol. 9 No. 1, pp. 29 – 43.
- Prinz, Jesse J. 2007. *The Emotional Construction of Morals*. New York, NY: Oxford University Press.
- Prinz, Jesse J. 2008a. "Resisting the Linguistic Analogy: A Commentary on Hauser, Young, and Cushman," in Walter Sinnott-Armstrong (ed.) *Moral Psychology Vol. 2. The Cognitive Science of Morality: Intuition and Diversity*, pp. 157 – 170. Cambridge, MA: The MIT Press.
- Prinz, Jesse J. 2008b. "Is Morality Innate?" in Walter Sinnott-Armstrong (ed.) *Moral Psychology Vol. 1. The Evolution of Morality: Adaptations and Innateness*, pp. 367 – 406. Cambridge, MA: The MIT Press.
- Prinz, Jesse J. 2008c. "Acquired Moral Truths," in *Philosophy and Phenomenological Research*, Vol. 77 No. 1, pp. 219 – 227.
- Prinz, Jesse J. 2012. *Beyond Human Nature: How Culture and Experience Shape our Lives*. New York, NY: Allen Lane.
- Rauscher, Frederick. 1997. "How a Kantian Can Accept Evolutionary Metaethics," in *Biology and Philosophy*, 12, pp. 303 – 326.
- Richerson, Peter J.; and Boyd, Robert. 2005. *Not by Genes Alone: How Culture Transformed Human Evolution*. Chicago: The University of Chicago Press.
- Ridley, Matt. 2003. *Nature via Nurture: Genes, Experience and What Makes us Human*. London: Fourth Estate.

- Rizzolatti, Giacomo; and Fogassi, Leonardo. 2007. "Mirror Neurons and Social Cognition," in R. I. M. Dunbar and Louise Barrett (eds.) *The Oxford Handbook of Evolutionary Psychology*, pp. 179 – 196. New York, NY: Oxford University Press.
- Rose, Steven. 1980. "'It's Only Human Nature': The Sociobiologist's Fairyland," in Ashley Montagu (ed.) *Sociobiology Examined*, pp. 158 – 170. New York, NY: Oxford University Press.
- Roskies, Adina. 2003. "Are Ethical Judgments Intrinsically Motivational? Lessons from 'Acquired Sociopathy'," in *Philosophical Psychology*, Vol. 16 No. 1, pp. 51 – 66.
- Rottschaefer, William H.; and Martinsen, David. 1990. "Really Taking Darwin Seriously: An Alternative to Michael Ruse's Darwinian Metaethics," in *Biology and Philosophy*, 5, pp. 149 – 173.
- Ruse, Michael. 1990. "Evolutionary Ethics and the Search for Predecessors: Kant, Hume, and All the Way Back to Aristotle?" in *Social Philosophy and Policy* Vol. 8, Issue 1, pp. 59 – 85.
- Ruse, Michael. 2006. "Is Darwinian Metaethics Possible (And If It Is, Is It Well Taken)?" in Giovanni Boniolo and Gabriele de Anna (eds.) *Evolutionary Ethics and Contemporary Biology*, pp. 13 – 26. New York, NY: Cambridge University Press.
- Ruse, Michael; and Wilson, Edward O. 1986. "Moral Philosophy as Applied Science," in *Philosophy,* Vol. 61 No. 236, pp. 173 – 192.
- Saltzstein, Herbert D; and Kasachkoff, Tziporah. 2004. "Haidt's Moral Intuitionist Theory: A Psychological and Philosophical Critique," in *Review of General Psychology*, Vol. 8 No.4, pp. 273 – 282.
- Scaltsas, Patricia. 1992. "Do Feminist Ethics Counter Feminist Aims?" in Eve Browning Cole and Susan Coultrap-McQuin (eds.) *Explorations in Feminist Ethics*, pp. 15 – 26. Bloomington: Indiana University Press.
- Slote, Michael. 2006. "Moral Sentimentalism and Moral Psychology," in David Copp (ed.) *The Oxford Handbook of Ethical Theory*, pp. 219 – 239. New York, NY: Oxford University Press.
- Slote, Michael. 2007. *The Ethics of Care and Empathy*. New York, NY: Routledge.
- Smith, Michael. 1994. *The Moral Problem*. London: Blackwell Publishing.
- Sober, Elliott, and Wilson, David Sloan. 1998. *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.
- Sterelny, Kim. 2001. "Niche Construction, Developmental Systems, and the Extended Replicator," in Susan Oyama, Paul E. Griffiths, and Russell D. Gray (eds.) *Cycles of Contingency: Developmental Systems and Evolution*, pp. 333 – 350. Cambridge, MA: The MIT Press.
- Street, Sharon. 2006. "A Darwinian Dilemma for Realist Theories of Value," in *Philosophical Studies*, 127, pp. 109 – 166.
- Sunstein, Cass R. 2005. "Moral Heuristics," in *Behavioral and Brain Sciences*, 28, pp. 531 – 573.
- Tavris, Carol. 1992. *The Mismeasure of Woman*. New York, NY: Touchstone.
- Thompson, Michael. 1995. "The Representation of Life," in Rosalind Hursthouse, Gavin Lawrence, and Warren Quinn (eds.) *Virtues and Reasons: Philippa Foot and Moral Theory*, pp. 247 – 296. New York, NY: Oxford University Press.
- Tooby, John; and Cosmides, Leda. 1992. "The Psychological Foundations of Culture," in Jerome H. Barkow, Leda Cosmides , and John Tooby (eds.) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, pp. 19 – 136. New York, NY: Oxford University Press.

- Trivers, Robert L. 1971. "The Evolution of Reciprocal Altruism," in *The Quarterly Review of Biology*, Vol. 46 No. 1, pp. 35 – 57.
- Waddington, Conrad H. 1942. "Canalization of Development and the Inheritance of Acquired Characters," in *Nature*, Vol. 150 No. 3811, pp. 563 – 565.
- Williams, Bernard. 1981. "Internal and External Reasons," in *Moral Luck*, pp. 101 – 113. Cambridge: Cambridge University Press.
- Williams, Bernard. 2002. *Truth and Truthfulness: An Essay in Genealogy*. Princeton, NJ: Princeton University Press.
- Wilson, Edward O. 1978. *On Human Nature*. Cambridge, MA: Harvard University Press.
- Wilson, Edward O. 1998. *Consilience: The Unity of Knowledge*. London: Abacus.
- Wrangham, Richard. 2009. *Catching Fire: How Cooking Made Us Human.* London: Profile Books.
- Zangwill, Nick. 2003. "Externalist Moral Motivation," in *American Philosophical Quarterly*, Vol. 40 No. 2, pp. 143 – 154.