Evolution and Moral Ecology

Timothy Dean

A thesis in fulfilment of the requirements for the degree of

Doctor of Philosophy

UNSW

THE UNIVERSITY OF NEW SOUTH WALES

School of Humanities and Languages

Department of Philosophy

University of New South Wales

September 2014

Acknowledgements

Only one name is attached to a thesis, but it takes many people to bring it into being. So I would like to offer my heartfelt thanks to everyone who has advised and guided me in the production of this work, particularly to my supervisor, Stephen Cohen, and co-supervisor Robert Brooks, along with my former co-supervisors David McKnight and Peter Slezak. Your guidance has been invaluable and your tolerance of a discipline-bending thesis has been nothing short of heroic. I would also like to offer special thanks to John Wilkins, who suggested the term "moral ecology" to capture the thrust of my core idea even while it was still taking shape. There are also many who have aided me through fruitful conversation and/or debate on a wide range of related topics, including Kim Sterelny, Paul Griffiths, Mike Manefield, Ben Fraser, Matt Hammerton, Tibor Molnar, Alkistis Elliott-Graves and my many postgraduate friends and colleagues at the University of New South Wales, University of Sydney and Australian National University. A special thanks goes to David Kidd, who has long been a dear friend, patient sounding board and an insightful advisor. I would also like to thank my friends and family, who have been nothing but supportive throughout this journey. I'd particularly like to acknowledge my mother, Diana, and father, Alan, for nurturing my curiosity since childhood with a rich environment of enquiry, wonder and love. A special mention goes to Stuart Maynard and Maynard's Café in Newtown, where most of the words in this thesis were written. And a singular thanks goes to my brilliant and beautiful partner, Signe, not only for undertaking the mammoth task of editing this thesis, but for many years of insightful advice and loving support. This thesis could not have been completed without you.

Table of Contents

Page

Chapter 5: The Moral Domain

Chapter 6: Moral Functionalism

Chapter 7: Cooperative Complexity

Chapter 8: Moral Ecology

Chapter 9: Moral Adaptation

Chapter 10: Adaptive Complexity

Chapter 11: Moral Niche Construction

Chapter 12: Moral Ecology In Action

Chapter 13: Moral Politics

Chapter 14: Evolution of a Complex Mind

Chapter 15: Evolution and Moral Diversity

Chapter 16: Moral Ecology and Diversity

Chapter 17: Conclusion

Chapter 1: Introduction

> Scientists and humanists should consider together the possibility that the time has come for ethics to be removed temporarily from the hands of philosophers and biologized.
> - Edward O. Wilson

1.0: The light of evolution

There is a long and somewhat chequered history of attempts to draw ethical insights from the theory of evolution, dating back to the progenitor of evolution himself, Charles Darwin. This thesis represents another such attempt, albeit one with a more modest objective compared to many that have come before. Rather than drawing on the theory of evolution to glean normative recommendations, as did the social Darwinists of the late 19th and early 20th century, or arguing that evolution undermines morality altogether, as have more recent metaethicists such as Richard Joyce (2006) and Sharon Street (2006), I aim to use evolution as a tool to help shed light on one particular moral phenomenon: moral diversity.

Diversity is often rather uninteresting, in biology as elsewhere. This is particularly the case if the diversity is merely a consequence of random variation. In fact, diversity used to be considered a nuisance by many biologists when it came to finding order in the variety of forms that existed within a species. The primary concern of early taxonomists was to divvy up the natural world into species by carefully scrutinising the phenotypic similarities between individuals. Yet it was observed that some creatures had slightly longer claws, some exhibited duller plumage, and some emitted a higher pitched cry. Even so, it is perhaps not surprising that an emphasis on *similarities* between individuals had the effect of obscuring their *differences*, which were often seen as little more than noise.

Yet individual differences often are quite interesting, in biology as elsewhere. And it was through Charles Darwin's theory of evolution by natural selection that the significance of difference came into sharp relief (Darwin, 1872). Firstly, it was this difference – this very noise – that was the raw material for natural selection. But this does not exhaust what is interesting about diversity, nor the significance of evolution in understanding individual differences. For, sometimes, diversity is not only the product of random variation, but is the product of natural selection itself. Sometimes the variation in claws, plumage or vocalisation is itself shaped, and maintained, by natural selection. The variation in beak size of Darwin's finches, or the existence of two sexes in many species, are examples of variation within a species that are maintained by evolutionary forces.

Likewise, I suggest, with the phenomenon of moral diversity. That there exists disagreement over a great many issues of moral importance is a trivial observation. That much of this disagreement is due to stochastic forces, such as differing circumstances, experience or upbringing is also unremarkable. But that some of this disagreement might exist for a very good evolutionary reason is a bolder claim.

The focus of this thesis is the question of why moral diversity exists, and why it exhibits the patterns it does. I will be arguing that at least some moral diversity has its roots in the complexities of cooperative interaction and in our nature as evolved social animals. It is worth stressing from the outset that my aim is not to provide an exhaustive account of why moral diversity exists, but simply to shed light on some of the evolutionary forces that have contributed to moral diversity. It is my hope that a better understanding of the forces that have contributed to moral diversity in the world – particularly evolutionary forces – might show that moral diversity is often more interesting, and perhaps more benign, than many philosophers appear to think. The implications of this evolutionary perspective on moral diversity might also prove useful in dealing with a range of hotly contested metaethical issues, such as the conflict between moral realism and anti-realism, and between objectivism and relativism, among others.

1.1: Morality outside-in

Central to this thesis is the theory of evolution, which has proven a potent unifying force in biology; as Theodosius Dobzhansky wrote in 1973, "nothing in biology makes sense except in the light of evolution". Dobzhansky pointed out that without evolution to elucidate the mechanisms that direct how organisms are shaped, the staggering diversity of forms of life presents us with a deep mystery: why is it that there exist so many different organisms, many with redundant and overlapping features or habitats? Furthermore: why so many with clumsy and sub-optimal design? And why so many that occupy precariously tight environmental niches?

Likewise, without evolution, the startling biochemical similarity among all life forms also seems incredible. After all, virtually the whole gamut of life, from the pseudo-living viruses through to the titanic form of the blue whale, are based on the same underlying nucleic acid chemistry in the form of ribonucleic acid (RNA) and deoxyribonucleic acid (DNA). They also share many other metabolic similarities, such as the ubiquitous energy transport molecule adenosine triphosphate (ATP). The very power of evolution stems from its ability to provide a single broad framework that can weave together these disparate forms and elucidate both why there exists such diversity and also so many similarities. It could be said that while other biological disciplines seek to answer "what" and "how" questions, evolution lends the crucially important "why" (Mayr, 1961).

Yet evolution can also prove a unifying force well beyond the disciplinary bounds of biology. It can extend its explanatory power into many aspects of human behaviour, not least our capacity to reckon right from wrong and behave accordingly. Where moral psychology can help map our moral responses to different situations, and anthropology can catalogue the moral mores of various cultures around the world and throughout history, evolution can serve as an overarching framework that can explain not only why these responses and norms come to take the diverse forms they do, but why there exist such similarities among them as well.

Already, evolution has proven a most corrosive force to the once prevailing notion in Western thought that humans are somehow distinct from the rest of the works of nature. In some ways, evolution has been a great leveller; *Homo sapiens* is really no more biologically different from the other animals than a whistled tune is different from a symphony. While there can be no doubt that humans exhibit an unparalleled complexity of cognition and behaviour, there remains far more that we have in common with the rest of life than that which makes us distinct. Where once humanity looked up to the heavens for the story of its creation, evolution has brought our gaze firmly back down to earth. I suggest evolution might likewise ground some of our moral thinking.

Ultimately, humans are social animals, albeit startlingly complex examples of such. And it is in viewing the vast complexities of human behaviour through the lens of this simple fact that we can gain some insight into some stubbornly curious moral phenomena. In this thesis I intend to view morality not as an abstract system of unconditional norms – as did, say, Immanuel Kant (1964) – nor merely as a form of discourse – as do many metaethicists – but as a *natural phenomenon*: a phenomenon that emerges from the observation of behaviours and interactions among social animals, such as us. Just as an evolutionary biologist or behavioural ecologist might seek to explain the social behaviour of the Satin Bowerbird, or the Sepia cuttlefish, or the crane fly, in terms of the dynamics of the pressures placed upon it to find a mate in its particular environmental niche, we can employ similar tools to study human social behaviour. Behaviour also has the virtue of being observable. So too are many of the environmental forces that influence how successful a behaviour is at satisfying the interests – whatever they might be – of the agent.

This means the perspective I take on morality in this thesis is somewhat different from that taken in a great many ethical and metaethical texts. Look at the world through the lens of metaethics and you see a landscape populated with agents prone to issuing statements of approbation or disapprobation about the actions of others. They might provide reasons and justifications for their judgements, perhaps citing facts, intuitions or sentiments – or some combination of all three – in their defence. Yet, look at the world through the lens of biology and you see a landscape populated by self-interested organisms wandering about, bumping into each other, and occasionally saying "sorry." These organisms also tend to create, spread and conform to behavioural norms that often prevent such "bumps" in the future.

The former view of morality tends to begin with moral language as its starting point, hoping to reveal something of the nature of morality by analysing the meaning of moral terms. As Don Loeb (2008) puts it, it is somewhat akin to attempting to "pull a metaphysical rabbit out of a semantic hat". The idea is that if, for example, "murder is wrong" is found to be a statement of fact, akin to saying "the sky is blue," then this raises the prospect that there might be a domain of uniquely moral facts that make such a statement either true or false. This, in turn, might suggest that ethics is akin to a science, whereby the discovery of the pertinent moral facts might aid us in arriving at the "correct" answers to moral questions, thus resolving many moral disputes. Alternatively, if it turns out that these moral facts do not exist, then that might call into question the veracity of our entire moral discourse. I call this perspective on moral discourse – in a somewhat affectionately tongue-in-cheek manner – the "inside-out" view of morality, as I will elaborate in chapter 4. It seeks to tell a story about morality starting with the inner world of introspection and working outwards from there primarily via language to the end point of behaviour.

There appears to be abundant interest amongst ethicists and metaethicists today in questions about whether or not moral sentences are truth-apt. However interesting such questions might be, they are of secondary importance when it comes to understanding and explicating our moral experience and behaviour. After all, it might turn out that our moral discourse is thoroughly confused, perhaps muddling a variety of concepts, or it might be used in different ways by different people, as suggested by Walter Sinnott-Armstrong (2009). If this is the case, it makes it rather difficult to derive any reliable conclusions about the nature of morality purely by reflecting on the content of our moral intuitions and discourse.

As such, this thesis takes an alternative perspective, which I dub the "outside-in" view on morality, as I will elaborate in chapter 4. This perspective begins not with language but with *behaviour*. For morality, as it is practiced, appears not only to be about moral utterances, beliefs and reasons about right and wrong, but to be importantly about how we social creatures behave, particularly how we behave towards one another. Morality appears to be a tool to guide how we act, how we respond to others' actions and how we urge others to act, particularly in social situations. It seems to inspire the "sorry" mentioned above when two or more organisms – or their interests – collide. The outside-in view then works its way inwards to look at the psychological processes that guide our behaviour and language use, rather than the other way around. As such, when seeking an explanation for an individual's moral behaviour, it is less inclined to take on face value what agents *say* are the justifications for their moral attitudes and judgements, and instead prioritises what they *do*. The outside-in perspective is not a new creation. In fact, it is the predominant approach used in a variety of disciplines to understand complex human behaviours, such as in sociology, social psychology and behavioural ecology. It is just a rather rare perspective to be employed by a philosopher.

The outside-in approach I take in this thesis is primarily descriptive and explanatory in nature and, as such, it places a premium on empirical evidence. This evidence is specifically geared towards exploring the causes of moral phenomena, in particular those psychological mechanisms that give rise to what we typically call moral judgements and moral behaviour. Yet this is a thesis of philosophy rather than science, belonging to that brand of philosophy which seeks to synthesise findings from a wide range of disciplines and apply them to philosophical questions. In doing so, this thesis adds to a growing chorus of philosophers, psychologists and scientists who believe evolution, biology and science are crucial tools in helping us to understand the phenomenon of morality, dating back to John Dewey (1922) and more recently Michael Ruse & Edward O. Wilson (1986), Joshua Greene (2002), Owen Flanagan, Hagop Sarkissian and David Wong (2008), Jonathan Haidt and Selin Kesebir (2010) and Phillip Kitcher (2011).

1.2: Explaining diversity

The core explanandum of this thesis is the phenomenon of moral diversity. This is, broadly speaking, the profound variation in moral attitudes and norms that appears to exist among individuals and among cultures, both throughout the world and throughout history. As I will discuss in the next chapter, there exists a tension between the apparent objectivity of our moral attitudes and the existence of widespread moral diversity and disagreement in the world. If our morality is grounded in objective facts, as many believe it to be, then presumably moral diversity ought to dissolve under the light of sufficient rational scrutiny. Yet, despite philosophers' best efforts at providing such scrutiny, diversity and disagreement persist.

In light of this, I am sympathetic to the view offered by many philosophers, such as John Mackie (1977), David Wong (1991), Richard Joyce (2001), John Doris and Alexandra Plakias (2008) and Phillip Kitcher (2011), who are sceptical about the existence of objective facts that ground our moral attitudes and judgements in truth. Rather, the existence of widespread, and apparently intractable, moral disagreement is good evidence that there is no objective fact of the matter that can resolve many moral disputes.

If this scepticism is warranted, then, as Kitcher (ibid.) points out, the notion of "moral progress" – the apparently progressive change in moral codes throughout history – becomes something of a mystery. For example, how is it that slavery or capital punishment can be considered to be *de rigueur* at one point in history and then come to be considered morally inexcusable without recourse to the claim that people have become aware of new facts concerning the impermissibility of slavery? If morality were more like science, one might indeed expect our attitudes about the world to increase in veracity over time, and this might result in something that resembles a progressive change in moral norms as they conform to our greater understanding of the facts. However, if such facts do not exist, then there is a risk that apparent moral progress dissolves into being "mere change", as Kitcher puts it.

It is my hope and expectation that an evolutionarily-informed outside-in perspective on morality might shed some light on this problem of reconciling the two apparently contradictory facets of moral experience, and can help explain the apparent progressiveness of moral change as a kind of "moral evolution." To begin, I will offer an explanatory story about the origins and causes and dynamics of moral diversity. Following is a brief outline of how this story will progress, indicating where further elaboration will take place in later chapters.

1.3: Moral functionalism

The outside-in view lends itself to a functionalist definition of morality, i.e. emphasising what morality *does* rather than what morality *is*. This functionalist definition, which I offer in chapter 6, paints a picture of morality as being a code of conduct that has historically served (and may still serve) the function of solving the problems of social living that disrupt cooperative interaction within groups, thus facilitating prosocial and cooperative behaviour[1]. Its ultimate function has effectively been about enabling individuals to co-exist and cooperate in such a way that they could pursue their interests – whatever they might be – without disrupting social harmony by compromising the interests of others. On a more proximate level, the functions of morality have included sanctioning self-interested behaviour, punishing free-riding, fostering group cohesion and identity, and enabling coordinated activity. Morality effectively enabled greater levels of cooperation within groups, and competition between groups, thus aiding the individuals within those groups to advance their interests – including their reproductive interests – more effectively than if they lived a less social and less cooperative existence.

It is important to stress that such a functionalist definition of morality is orthogonal to the presumed origins or justification of a moral system as articulated by its followers, such as whether they consider it to be strictures imposed by a deity or being based on respect for the inalienable autonomy of others, or some other foundation, natural or otherwise. One virtue of an outside-in functionalist definition is that it is not contingent upon what the adherents to a particular moral code think or say about its origins or justification. Rather it is based on observing the role that the moral code plays in steering their behaviour within their social group, which is largely an observable phenomenon.

This functionalist view sees moral norms as being fundamentally about guiding behaviour, particularly guiding behaviour towards satisfying the function of morality of facilitating social living and cooperation within groups. This is not unlike the perspective taken on social norms within the social sciences, particularly since the mid-20th century, a view dating back to the likes of Herbert Spencer (1883), Emile Durkheim (1895) and more recently Talcott Parsons (1937) and Robert K. Merton (1968), among others. This is inherently a teleological approach to function: moral norms appear to be designed to fulfil some goal and are judged on how well – or how "efficiently" in sociological parlance – they satisfy that goal. Norms tend to have a "problem background," which defines the

[1] Terms like "function" and "self-interest" will be more carefully defined in later chapters.

parameters of the problem to be solved by the norm, and give an indication of how to judge its effectiveness. Moral norms thus nest within this framework as a kind of social norm that specifically serves the function of promoting behavioural strategies that seek ultimately to promote prosocial and cooperative behaviour. The problem background that faces moral norms is constituted by the problems of social living, such as how to encourage group cohesion, the threat of self-interested behaviour destabilising social harmony or the problem of preventing free-riders. However, there are some crucial differences between the sociological and the evolutionary functionalist view that I espouse that will be elaborated in more detail in chapter 6.

Accepting a functionalist definition of morality raises a key question: if morality historically served the function of facilitating social living amongst self-interested individuals within a group, then what behavioural norms best promoted this end? As I will discuss in detail in chapter 7, the complex dynamics of social and cooperative interaction show that there is no easy answer to this question, and it is in untangling why this is the case that I suggest we can gain insight into the phenomenon of moral diversity. For it appears to be the case that, particularly when it comes to the many social and cooperative interactions that are modelled by game theory, there is no single behavioural strategy – or moral norm that promotes a particular behavioural strategy – that will consistently solve the problems of social living and promote optimal levels of cooperation in every environment. Furthermore, there appears to be no one *set* of strategies – or *set* of norms – that will solve these problems and promote optimal levels of cooperation in every environment. Instead, the success of any particular behavioural strategy or norm depends on the environment in which the behavioural strategy or norm is employed. Some norms might prove successful in multiple environments. Other norms might be highly successful in a particular environment only to fail dreadfully in another. For example, a norm that promotes trust in strangers might prove highly productive for individuals within an environment of abundant resources populated by willing cooperators. However, that very same norm might be disastrous for an individual to employ in an environment with scarce resources populated by individuals willing to defect for their own advantage (Bicchieri, Duffy, & Tolle, 2004). A norm might also be highly successful at one time only to become less effective at a later time, particularly if that norm has changed the very dynamics of the social environment in which it exists (Axelrod, 1997).

1.4: Moral ecology

I call the product of the dynamic and environment-dependent nature of moral norms "moral ecology,"[2] a term that will feature prominently throughout this thesis. The word "ecology" is intended in a metaphorical sense to invoke the biological concept of ecology, which describes the complex interactions that take place among a variety of organisms within an ecosystem. From an ecological perspective, the success of an individual organism depends on the environment in which it exists, with the environment including features of the physical and biological world in which it exists along with features of its own population. Within ecosystems, many organisms occupy or carve out their own niche in which they are successful, and there is no organism that is successful in every niche. What emerges is a community of different organisms, morphs, phenotypes and behavioural strategies that co-exist in a dynamic equilibrium over time.

Likewise, moral norms find their "niche" in those environments where they prove successful, although no norm is successful in every niche. Moral norms also change the very social – and sometimes physical – environment in which they operate, eventually forming into a dynamic equilibrium via a process of cultural evolution. Moral norms can also be seen to be competing with each other, and entire moral systems competing with neighbouring systems, in terms of attracting adherents and better facilitating cooperation and productivity. Groups of norms can be invaded by new norms, and individuals adhering to one set of norms can be invaded by individuals adhering to another.

One intriguing finding that I will explore in some detail in chapter 7 is that in many environments it often takes a pluralism of norms – for example, some promoting trusting behaviour and some promoting suspicious or punishing behaviour – in order to find a relatively stable equilibrium (Lomborg, 1996).

Moral ecology is also intended to be reminiscent of behavioural ecology, which studies the behavioural strategies that organisms employ in order to advance their evolutionary fitness. Both behavioural ecology and moral ecology see organisms as attempting to satisfy their interests, and as employing behavioural strategies in an attempt to advance those interests. The chief distinction is that in behavioural ecology these interests are fundamentally biological interests, whereby a successful strategy is one that promotes an

[2] I am aware that "moral ecology" is not an entirely novel term, and has been employed in the social sciences, albeit in a limited way, to mean something somewhat different from my stipulation (Hertzke, 1990). I hope my rendering is sufficiently clear that there will be no conflation of the terms.

individual's reproductive fitness. In contrast, moral ecology takes a more restrictive view, whereby a successful strategy is one that solves a problem of social living, thereby facilitating greater social and cooperative behaviour. The reason for this distinction is that behavioural ecology considers successful many socially disruptive strategies that would be regarded as either immoral or amoral, such as behaviours that are motivated by self-interest that might advance the interests of an individual at the expense of the interests of others, or at the cost of diminishing cooperation within a social group. Amoral behaviours include those that address prudential concerns that have no impact on individuals within a social group, such as solitary hunting strategies or heuristics used to avoid predators. It is important that any account of morality excludes such behaviours from its repertoire. As such, moral ecology nests within behavioural ecology, but only addresses a subset of the phenomena that are the concern of behavioural ecology, namely those that serve the function of morality, as defined above.

The notion of moral ecology also suggests that no population will be immune to perturbation or disruption. In fact, all things being equal, populations that enjoy high levels of cooperation will likely be more vulnerable to disruption than populations that consist of less cooperative strategies. Yet individuals and populations with low levels of cooperation will suffer a competitive disadvantage against individuals and populations that enjoy high levels of cooperation. Thus we would expect dynamism to be a common feature when looking at moral norms and behaviour among and within cultures.

Moral ecology is intended to capture the highly dynamic nature of social and cooperative interaction, and informs the problem background of promoting prosocial and cooperative behaviour, which has been the core function of morality. It is in understanding these dynamics that we can understand not only why moral diversity exists in the world, but why the specific patterns of diversity we observe come to be so.

1.5: Evolution and moral psychology

As I will discuss in chapter 14, the dynamics of moral ecology that have contributed to some moral diversity in the world also may have influenced the evolution of our minds. The tremendous benefits of cooperation born from social living appear to have provided a strong selective pressure in our hominin ancestors driving the evolution of psychological mechanisms that promote such social and cooperative behaviour (Humphrey, 1976; Sterelny, 2003). Just some proposed psychological mechanisms that contribute to our social and cooperative behaviour include the "moral" emotions, such as empathy, guilt and righteous anger (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Haidt, 2001; Huebner, Dwyer, & Hauser, 2009); heuristics that promote incest avoidance (Lieberman, Tooby, & Cosmides, 2003; Lieberman, 2008); social imitation and in-group conformity (McElreath, Boyd, & Richerson, 2003); and a tendency to create and adhere to behavioural norms (Sripada & Stich, 2005). In concert, these mechanisms appear to have been capable of overcoming the substantial evolutionary hurdles to encouraging prosocial behaviour and cooperation in unrelated groups of individuals (Kitcher, 2011).

Yet, while there has been tremendous progress over the past decade in revealing how our minds have been shaped by evolution, relatively little attention has been given to the notion that individual differences in our psychology might also be the product of evolutionary forces. Furthermore, some of this evolved variation in our psychological traits and dispositions might contribute to at least some of the moral diversity observed within cultures. Most evolutionary psychology also concerns itself with uncovering the universals of human nature (G. F. Miller, 2001; Sidanius & Kurzban, 2003; Tooby & Cosmides, 1990). And most moral psychology studies have sought universal patterns that underlie moral judgements formed by normal functioning individuals within a particular environment (Greene & Haidt, 2002; Greene et al., 2009; Hauser, 2006). However, relatively little evolutionary or moral psychology literature has addressed the influence of evolution on individual differences in psychology and cognitive function – although there are rare exceptions (Bateson, 2004; Buss, 2009; D. S. Wilson, 1994) – and even less on the influence that individual differences have on moral diversity and moral disagreement.

Yet there is compelling evidence that variations in our psychological makeup contribute to variations in our moral attitudes and judgements. As I will discuss in chapter 13, much of this evidence comes not from moral psychology but political psychology. This is a discipline that seeks to find the psychological factors that influence our political attitudes and behaviour. Yet political attitudes bear significant enough similarity to moral attitudes, particularly when they concern issues like equality, reciprocity and harm, that the findings in political psychology may have bearing on our understanding of moral psychology. For example, studies in political psychology have found that psychological traits such as personality type, sensitivity to fear and perception of threat, and tolerance of ambiguity, among others, correlate with many political attitudes (Frenkel-Brunswik, 1948; Jost et al., 2007; Mondak, 2010). Furthermore, it appears that many political attitudes are also heritable to at least some degree, suggesting some of the variation in political attitudes within a population is due to variation in genes (Alford, Funk, & Hibbing, 2005). Given that many psychological traits are also heritable (Bouchard & McGue, 2003; McCrae & Costa Jr., 2003), this raises the prospect that genetic variation might influence our political and moral judgements. This, in turn, raises the prospect that there might be an adaptive story to tell about the existence of such genetic variation and its ability to produce adaptive social behaviour.

I will present two theses that draw a link between evolved psychological variation and variation in moral/political attitudes: the weak thesis and the strong thesis. Both argue that evolved variation in psychological traits contributes to variation in moral/political attitudes, but they differ on whether such variation in a by-product of evolutionary forces, or whether the variation in moral/political attitudes was itself adaptive. Both suggest that it was not possible for evolution to gravitate towards a single set of psychological traits or behavioural dispositions that could produce adaptive social behaviour in every environment. This is because of the highly complex and heterogeneous social environment in which we evolved – and environment that exhibits many of the strategic interactions and complex dynamics discussed in chapter 7. As such *Homo sapiens* evolved a diverse and highly plastic psychology that can respond to a wide range of environmental conditions. This includes not only a range of behavioural predispositions and heuristics (Cosmides & Tooby, 1997) and our signature unparalleled cognitive plasticity (Godfrey-Smith, 1998), but also a powerful tendency to create and adhere to norms (Sripada & Stich, 2005) and to produce culture (R. Boyd & Richerson, 1985). Together these aid in producing adaptive behaviour, particularly in complex social environments.

This strong thesis in particular argues that variation in psychological function is far from just random noise or due to genetic drift. In fact, the variation itself appears to be adaptive. And this would not be the only instance of adaptive variation that persists within a population (D. S. Wilson, 1994). One particular evolutionary mechanism that appears to be responsible for maintaining psychological variation is balancing selection, and negative frequency-dependent selection in particular, whereby the fitness of a trait increases as its frequency in the population decreases. For example, a trait that encourages bold or aggressive behaviour can increase in fitness the more rare it is in a population, yet the same trait can decrease in fitness within a population filled with other bold or aggressive individuals (Maynard Smith, 1982). A classic concrete example of negative frequency-dependent selection in action is the balanced sex ratio found in many species, including our own. I will argue in chapter 14 that much of the variation in human psychology may be maintained by balancing selection, along with other mechanisms such as bet hedging, and that this variation contributes to moral diversity.

There is also evidence that most people do not tend to see their moral judgements as being subjective expressions of their sentiments or acknowledgement of social conventions, but rather as being perceptions of objective facts about the world (Turiel, 1983). As such, it appears that our moral psychology lends itself to a form of "projectivism," whereby we "project" our sentiments on to the world, a view that dates back to David Hume and has more recently been expounded by Simon Blackburn (1993). As a result, it is natural for many people to interpret their moral utterances as being statements of fact made true or false by recourse to objective facts – a metaethical position known as cognitivism, and one that lends itself to moral realism. However, if it is the case that our moral judgements are, in fact, expressions of our sentiments rather than statements of objective fact, an anti-realist interpretation might be the correct one (Greene, 2002). As I will discuss in the final chapter, this projectivist phenomenon might account for our tendency to view morality in an objectivist realist sense, but also helps explain why such a metaethical view is in error.

1.6: Moral ecology and ethics

Thus I argue that at least some moral diversity may exist in response to the dynamic forces of social interaction and the complexities of solving the problems of social living. On the one hand is the diversity of psychological traits that have evolved in this complex adaptive environment. On the other is the messy process of cultural innovation that has produced norms in response to the problems presented by the specific proximate environment in which that culture resides. And, as such, to the degree that our present and future social environments exhibit similar levels of complexity and dynamism, and to the degree that we still seek to advance the function of morality to solve the problems of social living, it might be the case that some moral diversity is still warranted. Thus, rather than moral diversity being merely emblematic of widespread error, it might exist for very good reasons, and may even be beneficial. This is one of the valuable implications of the theory of moral ecology if it is true.

In a sense, this idea is reminiscent of a comment made by John Mackie in his 1977 book *Ethics: Inventing Right and Wrong* when he conjectures that moral disagreement might be better explained in terms of "people's adherence to and participation in different ways of life" rather than the "hypothesis that they express perceptions, most of them seriously inadequate and badly distorted, of objective values" (Mackie, 1977). One might consider the notion of moral ecology as cashing out in ecological and game theoretic terms what might be meant by Mackie's term "ways of life", and how they impact moral attitudes and norms. In a similar vein, this thesis might be read as lending support to the naturalistic anti-realist account of morality, such as that proposed by the likes of John Mackie, Michael Ruse, Owen Flannagan or Joshua Greene.

As should already be apparent, this is a thesis about *morality*, yet there is a great deal of talk about *cooperation*. This is not to suggest that the terms are synonymous, nor that promoting cooperation is all there is to morality. However, through understanding the dynamics of cooperation and the ways in which our psychology has evolved to promote cooperation, we can gain considerable insight into a great deal of morality. A comprehensive descriptive account of the dynamics and challenges of promoting cooperation may not answer all questions in moral philosophy, yet many questions in ethics cannot be answered without a comprehensive descriptive account of the dynamics and challenges of promoting cooperation. In light of this, I will not attempt to argue definitively that morality is only about cooperation, but suggest that to the extent that the dynamics of cooperation are relevant to understanding moral diversity, then this thesis will also be relevant to understanding moral diversity. I will, however, endeavour to avoid conflating the two terms in this thesis, and will offer more detailed definitions of key terms such as "morality" and "cooperation" in chapter 5. The first step is to explore the "problem" of moral diversity in more detail, which will occupy the next chapter.

Chapter 2: Moral Diversity

> Everyone without exception believes his own native customs, and the religion he was brought up in, to be the best.
> - Herodotus

2.0: Diversity and disagreement

It is unlikely you would approve of the behaviour of a 5th century BCE Scythian war party should they drop by for tea. That is even should they arrive as guests rather than plunderers. Depending on their success in recent ventures, you might see the younger warriors encouraged to drink the blood of their first slain foe, and reprimanded by their elders if they refused. You might find them sorting and preparing the severed heads of their recently vanquished enemies for presentation to their king upon their return. They might be busily removing the scalps from those skulls and treating them to be worn as trophies, demonstrating their prowess at killing their foes to the approval of their peers. Should the Scythian king pop in as well, the warriors might pass around a skull that had been turned into a grisly vessel, out of which all the successful (i.e. blooded) warriors would drink a toast. Such behaviours, while common and morally sanctioned two and a half thousand years ago would likely illicit moral revulsion today, at least in polite company. And if it should happen that you failed to show moral revulsion towards these acts, your neighbours peering over the fence would likely express moral revulsion towards you too.

Granted this is a somewhat abstruse example, but it is readily apparent that differences in attitude about what constitutes right and wrong behaviour appear to be ubiquitous throughout the world and throughout history, even between individuals less distantly removed than citizens of modern liberal societies and 5th century BCE Scythian warriors.

We know some of the norms and mores of the ancient Scythians from the writings of the Greek historian, Herodotus, who was remarkable not only as an ambitious post-Homerian story teller but also for assembling a catalogue of the customs and practices of the ancient world. In the centuries prior to Herodotus's time the vast majority of individuals would have spent their entire lives within sight of their birthplace, and exclusively interacted with members of their own culture. However, the 5th century BCE was a time when many cultures began to rub shoulders, not least within the courts of great empires such as that of Darius the Great in Persia.

In *The Histories* Herodotus relates an illuminating tale of apparent moral disagreement from that very court, which occurred when Darius is said to have summoned the Greeks that happened to be present. He asked them for what sum of money would they eat the dead bodies of their fathers rather than enact their usual custom of cremating their dead. Their reply was as unequivocal as it was splashed with outrage: they would not commit such an act for any money in the world. At this point, Darius summoned members of the Callatiae tribe from India and asked them what it would take for them to burn the bodies of their dead parents rather than eat them. Their reply was equally firm: they are said to have uttered a cry of horror and pleaded that Darius never again mention such an abhorrent act. Even if it could be said that both the Greeks and the Callatiae might have agreed on an underlying principle of honouring one's parents, their views as to how one ought to show such respect appear to be thoroughly incommensurable. Herodotus follows this tale of the Greek interaction with the Callatiae with a famous, and tantalisingly relativist, observation that "one can see by this what custom can do, and Pindar, in my opinion, was right when he called it 'king of all'"[3] (Herodotus, 1996).

This passage from Herodotus appears to show more than just a trivial disagreement over arbitrary customs, such as manner of dress or certain rules of etiquette. It is a disagreement that elicits what most would call a distinctly *moral* response, in this case a response consisting of revulsion at the prospect of either cremating or consuming the dead, respectively, and likely a corresponding desire to prevent others from practicing the offending ritual. This is a response that seems unlikely to alter simply by the edict of some authority figure proclaiming that the old practice is now forbidden and the alternative is now obligatory; should Darius have declared that cremation was now the accepted practice, it seems unlikely the Indians' attitudes would have readily changed. Where there might be tolerance of some differing customs, morality is a more stubborn matter. When it comes to those practices that possess a moral dimension it appears as though many individuals are often unable to fathom, let alone forgive, the transgression of their own culture's moral values.

[3] As far as I can discern from unreliable sources, the poem from Pindar referred to by Herodotus says:

> Custom, the king of all
> of mortals and immortals,
> leads, justifying that which is most violent
> by its very powerful hand.

A difference in attitudes between individuals from remote cultures about what practices are prohibited or obligatory is only one form that moral diversity can take. In the following sections I will cite some examples of the phenomena of moral diversity in three broad spheres, each with its own distinctive dynamics and challenges when it comes to explicating them, and all of which will be considered through this thesis. In chapter 3 I will then go on to elaborate in more detail the philosophical problems that are raised by moral diversity, focusing on the tension between the existence of moral diversity and notion that morality is somehow objectively founded. In section 3.1 I will delve deeper into a range of responses to the problem of moral diversity from moral objectivists and realists, then in section 3.2 do the same from subjectivist and anti-realist positions. Finally, in section 3.3 I will offer an alternative perspective on the significance of moral diversity which will set the scene for my own thesis of moral ecology.

2.1: Inter-cultural diversity

One form that moral diversity can take is that of differences in the moral norms and attitudes that exist among different cultures. Such cultural and moral diversity is one prominent theme in Herodotus's *The Histories*, as it is in many contemporary anthropological and sociological texts. For example, the Finnish anthropologist and philosopher, Edward Westermarck, in *The Origin and Development of Moral Ideas* (1906) details countless comparative customs and laws of cultures ranging from tribes of Africa, Australia or the Polynesian islands, to more large-scale cultures such as the Aztecs or ancient Egyptians, to the more so-called "civilised" nations of ancient Rome and late 19th century Europe. The diversity of rules and practices documented by Westermarck is profound, just one example being the rules he cites in reference to death caused through self-defence:

> Among the Fjort, though a person who kills another in self-defence is exempt from punishment, he is expected to pay damages. Among the Hottentots self-defence is regarded as a mitigating circumstance, but not as an excuse in the full sense of the word. Among other peoples it is not considered at all. Among the ancient Teutons a person who committed homicide in self-defence had to pay *wer*; and in Germany such a person seems to have been subject to punishment still in the later Middle Ages. In England, in the thirteenth century, he was considered to deserve royal pardon, but he also needed it. (ibid.)

While there is a common underlying principle here – that killing in self-defence is an offence different in gravity from unprovoked killing – the actual norms vary considerably from one culture to the next. In some cultures, killing in self-defence is considered permissible, in others it elicits reprobation of some sort, but not to the same degree as unprovoked killing, while in others it was prohibited entirely. These differences appear to concern more than just arbitrary cultural habits, they appear to possess a clear moral character, such that it would not be unexpected to see outrage should one of these mores be transgressed. Yet they also vary considerably in their detail, to the extent that the application in one region would quite likely elicit outrage in individuals from a different culture.

One feature of these moral proclivities is that they tend to be “sticky,” in that even upon exposure to other cultures individuals seem resistant to readily changing their moral code or adopting the norms of another culture, at least not without external coercion. This very phenomenon was remarked upon by Herodotus:

> If anyone, no matter who, were given the opportunity of choosing from amongst all the nations in the world the beliefs which he thought best, he would inevitably, after careful consideration of their relative merits, choose those of his own country. Everyone without exception believes his own native customs, and the religion he was brought up in, to be the best … There is abundant evidence that this is the universal feeling about the ancient customs of one’s country. (Herodotus, 1996, Book III, section 38)

What is notable about this passage is that Herodotus is not just claiming that each of us adheres to the customs of our own culture by habit or ignorance, but that even after careful consideration we would still believe that our own culture is “the best”. It is not that we adhere to our own cultural practices and beliefs simply because we are presented with no alternatives, but rather there is something about our experience and enculturation that shapes our values and worldview such that our own culture appears not only preferable, but somehow *superior* to other cultures.

I call *inter-cultural moral diversity* a difference in moral attitudes or norms that exists between cultures, such that individuals adopting those norms would be prone to disagree over what constitutes right action in a particular situation.

Inter-cultural diversity has been a topic of great interest to philosophers, particularly to the extent that moral diversity tends to lead to moral disagreement. As I will discuss in detail in chapter 3, most individuals consider their moral views to be more than just subjective preferences. Yet, if moral views are grounded in objective facts, this presumably raises the prospect that disagreement can be resolved. What, then, to make of moral diversity? The multitude of moral views that have existed throughout the world have caused some philosophers to doubt the existence of objective moral facts, and to suggest that moral diversity is simply a result of there being many ways of living (Mackie, 1977; Westermarck, 1932; D. Wong, 1991). The challenge for the moral realist is to show that inter-cultural moral diversity is either a source of error, or to show that moral facts can allow for multiple conflicting views. The subjectivist and relativist, on the other hand, have to explain why it is that most people feel that their moral beliefs are objectively true, and even upon reflection of their own enculturation and circumstances, are often reluctant to change their views. As I will discuss in chapter 3, they also have to explain how a subjectivist and/or relativist account retains the binding imperative force that most people consider morality to have.

2.2: Temporal diversity

Historical and anthropological accounts, such as those of Herodotus and Westermarck, readily demonstrate that moral attitudes and norms do not just vary between cultures at a particular point in history, but also vary over time within a particular culture. Many cultures have seen their adopted norms change. As Philip Kitcher (2011) points out, there was a time when chattel slavery was accepted in many cultures around the world, and was readily endorsed even by post-Enlightenment European cultures such as the United States and Britain. Furthermore, punishments for transgressions against moral or legal rules used to be far more punitive than they typically are today. The *lex talionis* – or "an eye for an eye" – brand of strict retributionist justice has now largely passed out of fashion. Another example of a significant change in moral attitudes in many cultures compared to even half a century ago is the increase in the rights and freedoms of women.

Westermarck remarks that the history of moral ideas is one that sees noticeable "progression" in some areas, such as when it comes to the punishment meted out for adultery. Less "civilised" cultures appear more likely to resort to direct methods of enacting punishment, with the individual who suffered the grievance exacting retribution on the perceived violator, whereas more "civilised" societies tend to remove that responsibility from the hands of the injured party and place the power to punish in the hands of a third party, often bound by laws and regulations that apply broadly across all members of the society. The broad flow of this trend from individually-meted to state-based enforcement of moral reckoning forms the overarching theme of Westermark's *The Origin and Development of the Moral Ideas*. Similar points are raised by Jared Diamond in his comparison of the culture and mores of tribes in New Guinea and the moral systems and institutions of modern day Americans (Diamond, 2012).

I call *temporal moral diversity* a difference in moral attitudes or norms that exists within a culture between two discrete points of time, such that individuals adopting those norms would be prone to disagree over what constitutes right action in a particular situation. However, in many cases, it is sufficient to consider temporal moral diversity as being a special case of inter-cultural moral diversity. The cultures at two different points of time can simply be considered as two separate cultures – albeit likely with many similarities – that differ in regard to some of their moral norms or practices.

Where temporal moral diversity becomes interesting in its own regard is in considering the nature of moral progress. As Kitcher points out, temporal moral diversity presents something of a puzzle. One common view is that moral judgements are made true or false by recourse to objective facts – a position referred to as moral realism or moral objectivism, which I will discuss in more detail in chapter 3. If moral facts exist, then moral progress can simply be a case of us expanding our understanding of the world and employing reason to adjust our moral system in accord with our knowledge of new facts. However, there are good reasons to question the existence of moral facts, and to dispute whether moral judgements are grounded in non-moral facts, as I will discuss below. If moral judgements are not grounded in objective facts, then how do we distinguish genuine moral progress from “mere change”? Furthermore, Kitcher doubts whether many significant examples of progress are driven by a process of rational deliberation or the discovery of new facts. He remarks:

> The historical figures who figure in ethical transitions, the vast majority of them unidentifiable as individuals, do not start from some situation in which they lack ethical convictions, follow a process of reasoning or observe some facet of reality, and thereby arrive at a well-grounded belief in an ethical judgement… For the revisionary historical actors who stand out relatively clearly – Mary Wollstonecraft, John Woolman – what occurs is a *change* in ethical conviction: ethical beliefs transmitted within the society and shared by everybody around are rejected in favor of claims incompatible with them. (op. cit.)

As such, the existence of temporal moral diversity raises a number of questions that have significant ramifications on ethics and metaethics, such as on the nature of moral judgement, the truth or otherwise of moral realism and on the possibility of moral progress.

2.3: Intra-cultural diversity

A third dimension to moral diversity is difference and disagreement *within* cultures. This is where two or more individuals living within the same culture at the same time hold different attitudes about what constitutes right and wrong. Westermarck notes that it was just such a moral disagreement between himself and his peers that inspired his lengthy endeavour cataloguing the various moral predilections of different cultures throughout the world and through history in *The Origin and Development of Moral Ideas*:

> Its author was once discussing with some friends the point how far a bad man ought to be treated with kindness. The opinions were divided, and, in spite of much deliberation, unanimity could not be attained. It seemed strange that the disagreement should be so radical, and the question arose, Whence this diversity of opinion? Is it due to defective knowledge, or has it a merely sentimental origin? And the problem gradually expanded. Why do the moral ideas in general differ so greatly? And, on the other hand, why is there in many cases such a wide agreement? Nay, why are there any moral ideas at all? (Westermarck, 1906)

Moral disagreements such as this are a hallmark of the modern world perhaps even more so than the ancient (Bloomfield, 2001; Mackie, 1977; Stevenson, 1937; Tersman, 2006). In fact, disagreement over moral standards and norms appears to be a signature feature of contemporary liberal societies, concerning a wide range of issues, such as capital punishment, abortion, euthanasia, the moral and legal status of homosexuality and issues concerning fairness and equality when it comes to the distribution of wealth.

The challenge for philosophers is to explain how moral attitudes can come to vary even between individuals who have been enculturated in the same environment, presumably with access to the same facts about the world and the same reasoning processes to evaluate them. If it is the case that we absorb all our moral attitudes from our culture, then we might expect less variation in moral attitudes among individuals from the same culture than we observe today.

2.4: Defining moral diversity

I have outlined moral diversity in three broad spheres: inter-cultural; intra-cultural; and temporal. What remains is to define moral diversity in more specific terms, focusing on what these three types of moral diversity have in common, and to distinguish moral diversity from non-moral diversity. In elaborating any definition of moral diversity, much will hang on the definition of “moral.” I will have a great deal to say about what delineates the moral from the non-moral domain in chapter 5, but for now I will simply stipulate that moral diversity refers to situations where two or more individuals hold different views about what is morally permissible (or impermissible, obligatory, etc) in a certain situation. The disagreement in question could concern the permissibility or otherwise of a particular action, the proper application of a moral norm, which norms ought to be obeyed, justifications and reasons for a particular judgement, which virtues ought to be cultivated, or the general moral principles that one ought to adopt. Of course, to the degree that one's definition of "moral" or "morally permissible" varies, so too will the meaning of moral diversity. I will have more to say about the defining of morality in chapters 5 and 6.

Chapter 3: Moral Tension

> Probably to most students of Moral Philosophy there comes a time when they feel a vague sense of dissatisfaction with the whole subject.
> - H.A. Prichard

3.0: Remarkable diversity

Moral diversity would be thoroughly unremarkable were morality more like food. After all, food preferences and cuisines vary tremendously across the globe and throughout history, yet this trivial observation does not often lead people to question the validity of their own preferred dishes or the truth of others'. A great deal of diversity in food attitudes is plausibly accounted for by biological predisposition to prefer certain flavours (i.e. sweet and salty compared to sour and bitter), by variation in culture, including enculturated tastes and parental influence, along with environmental variables like available produce and local farming practices (Birch, 1999). Even if we might be prone to making the occasional objectivist or normatively suggestive utterances about food ("Vegemite is awful, you should not eat it!"), most of us will concede that we are really expressing subjective preferences: we do not disapprove of Vegemite because it really is objectively bad; rather we disapprove of it because we are not fond of the taste. And if our tastes changed, then we would likely approve of its consumption. While people are known to argue over their subjective preferences for different foods, there is little argument about there being "one true cuisine" in the way that some claim there is "one true morality" (Harman & Thompson, 1996; M. Smith, 1994).

However, it seems that moral preferences are typically considered by most people to be a fundamentally different kind of thing than food preferences. Differences in moral judgement or belief are typically not tolerated in the same way as disagreements over subjective preference. If someone asserts that "torturing innocents is wrong," they are generally not simply taken to be expressing a subjective preference, like they might assert that Vegemite is good or bad. Nor would we expect them to tolerate someone else practicing torture based on that individual's assertion that torture is permissible. If they were to encounter another individual who held different views on the permissibility of torture, we would expect them to disagree with each other and perhaps engage in moral argument over whose position is the correct one.

Moral diversity and disagreement have proven to be the subjects of interest to metaethicists particularly in the context of the challenge they present to the objectivist realist account of moral discourse. For, if it were the case that there were objective moral facts that underpin moral judgements and beliefs, then presumably an appeal to such facts would help resolve any ethical disputes one way or the other. And if these objective facts were at least moderately transparent to human enquiry, then we might expect a greater convergence of belief in moral matters as our understanding of the world improved, much as there is greater convergence in belief about the natural world as science has progressed.

Yet, as numerous thinkers have pointed out (Brink, 1984; Mackie, 1977; Smith, 1994; Wong, 1991, among others), while there are parallels between ethical and scientific enquiry, there are some important differences. In the sciences, disagreement itself does not imply there is no fact of the matter that can settle the dispute. If there is disagreement over whether the Earth orbits the Sun, or the Sun orbits the Earth, there are generally agreed upon objective facts (knowable to various degrees of certainty) that can conclusively resolve the disagreement one way or the other. Even in disputes where the facts are more elusive, such as facts about events from the distant past or about the nature of objects remote from us or entirely inaccessible to observation, there is generally agreement that such facts exist nonetheless, and we simply have not discovered them yet. And where such facts are reliably known, the disagreement tends to dissolve away, at least amongst rational and well informed interlocutors. To the extent that ethical enquiry parallels scientific enquiry, we might expect a similar result. Yet while ethical discourse might sometimes resemble that of the sciences, there appear to be some disagreements that persist despite our best efforts to resolve them. Some disputes appear to concern issues whereby no objective facts can possibly resolve them, and some disputes may even resist resolution between rational and well informed interlocutors. Some moral disagreements appear to be genuinely *intractable*.

Disagreements in the sciences, even if they appear intractable today, do not tend to present sticky metaphysical problems as do some particularly gnarly disagreements in ethics, such as the permissibility of torture or our moral obligations towards the needy. In some cases of moral disagreement, such as the one recounted by Westermarck above over how well a “bad man” ought to be treated, it is plausible to imagine a situation where all interlocutors are in complete agreement over the empirical facts of the matter yet remain in disagreement over the moral outcome. As such, the persistence of seemingly intractable moral disagreement in an increasingly well-informed world suggests that such convergence is lacking, a notion that has led some to question whether objectivist realism is the best way to characterise morality (Blackburn, 1993; Joyce, 2001; Mackie, 1977).

Thus we arrive at the conundrum that exists at the heart of metaethical enquiry: the tension between the apparent objectivity of moral discourse and the apparent diversity of moral attitudes and norms in the world. In the following sections I will look at a range of views on the nature of morality – chiefly what I call "inside-out" metaethical views – that cash out these two facets of our moral experience in a variety of ways. I separate these metaethical views into two broad categories depending on whether they are more impressed by the apparent objectivity of morality or by the existence of moral diversity. I then examine the tension between these two broad approaches and look at the challenge of developing an account of morality that can adequately accommodate, or at least account for, both facets of our moral experience.

3.1: Moral realism

The claim to objectivity inherent in moral discourse mentioned above is often cashed out in the form of three distinct but overlapping theses. The first is a semantic thesis that moral statements, such as "Mary morally ought to do Φ" or "Φ is wrong" are "truth-apt." This means they express propositions that can be found to be true or false – a thesis referred to as *cognitivism*. This thesis is not necessarily committed to moral utterances *only* expressing propositions; they can also contain an expressive or non-cognitive component. However, cognitivists generally maintain that moral utterances possess at least some propositional content, and that this content is important to the moral utterance. The second is the metaethical thesis of moral *objectivism*, which states that morality is somehow "out there," that it is not mind-dependent, or constituted or justified by our subjective attitudes. Rather, objectivists assert that moral beliefs and judgements, such as "Mary morally ought to do Φ" or "Φ is wrong", are either true or false by recourse to some objective features of the world. The third, related, thesis is moral *realism*, which states that some of these moral statements actually are true, and they are made so by objective features of the world.

While cognitivism, objectivism and realism are independent theses, and can exist in isolation or in various exotic combinations, they tend to cluster together, largely because this triumvirate appears to systematise our everyday usage of moral language (Greene, 2002). If it were possible to show that the three together give a plausible account of how we use moral terms, and that usage is ratified by a consistent metaphysical picture, then many metaethical questions would be a lot easier to solve. For the sake of brevity, I will henceforth refer to the conjunction of cognitivism, objectivism and realism simply as "moral realism" or just "realism," as advocated by "moral realists" or just "realists," throughout the remainder of this thesis.

Judith Jarvis Thomson offers one neat example of a realist position with her "Thesis of Moral Objectivity", which she defines in terms of it being "possible to find out about some moral sentences that they are true" (Harman & Thompson, 1996). Michael Smith offers another paradigmatic account:

> We seem to think moral questions have correct answers; that the correct answers are made correct by objective moral facts; that moral facts are wholly determined by circumstances; and that, by engaging in moral conversation and argument, we can discover what these objective moral facts determined by the circumstances are. (M. Smith, 1994)

The relevant facts that make moral sentences true or false have been variously articulated as being either natural facts, such as facts about pleasure or human welfare, such as defended by the so-called 'Cornell Realists' (R. N. Boyd, 1988; Brink, 1989), or some special kind of moral facts, perhaps denoting some non-natural property of things, as was argued by G. E. Moore (1903).

A fourth thesis that is worth mentioning is moral *absolutism*, which states the objective moral facts that make particular moral utterances true or false apply to each and every individual in like circumstances. Absolutism appears to capture something of the unconditional nature of moral utterances, such that saying "torture is wrong" is not typically considered to apply contingently given an individual's beliefs, desires or whether they actively subscribe to such a norm. One exemplar of moral absolutism is Immanuel Kant, who argued that morality binds us to act in certain ways simply by virtue of us being rational beings, just as it binds all like rational beings in like circumstances (Kant, 1785). As such, if torture is wrong for A in circumstances C, then it is wrong for B in circumstances C. A particularly strong brand of absolutism is articulated by David Wong (who goes on to reject it), which is committed to the notion that "there is a single true morality for all societies and times" (D. B. Wong, 2006). Moral absolutism is not as popular a thesis as moral objectivism and realism, but it is worth stressing that it reflects another way that our moral experience is often cashed out. As such, I will focus my attention on realism, with reference to absolutism where it is relevant.

One of the appealing features of moral realism, besides its ability to explain one core aspect of our moral experience, is that it provides firm foundation for our sense that morality is somehow binding irrespective of our subjective attitudes or desires; that morality demands “unconditional obligation”, as Jesse Prinz puts it (2007), or a conjunction of inescapability and authority that Richard Joyce terms “practical clout” (2001). As I will discuss in more detail below, if morality was constituted by our subjective attitudes and/or contingently binding depending on our desired ends, then morality would appear to lose a great deal of its “practical clout.” In such a subjectivist framework, one could simply express a different belief about what is right and wrong, or claim that behaving in a moral way does not serve their interests, or that they are simply uninterested in adhering to moral norms at all, and thus be excused from conforming with a particular moral norm. If, however, morality is grounded in objective facts, then an individual’s moral attitudes can be evaluated according to whether they are true or false irrespective of their subjective attitudes or desires.

This broad brush articulation of moral realism is not to say that there does not exist a startling constellation of different metaethical theories. Many of these theories reject one or more of the above theses that attempt to explain the objectivism implicit in much ordinary moral discourse. For example, non-cognitivists, such as C. L. Stevenson (1937 & 1950) and A. J. Ayer (1936), reject cognitivism from the outset by denying that moral utterances are statements of fact, favouring an analysis of them as purely expressions of approval or disapproval. However, while there may be an expressive component to many moral utterances, one needs to take a highly liberal, if not radically revisionist, approach to moral discourse in order to maintain that moral utterances are *only* expressive and lack a propositional component altogether.

Moral sceptics, on the other hand, such as John Mackie (1977) and Richard Joyce (2001), accept cognitivism (or, at least, reject radical non-cognitivism) but also reject realism, claiming that moral utterances do include statements of fact, but that the referenced moral facts do not exist. This implies a broad error theory that applies to all objectivist moral discourse; we talk as if there were moral facts, but our discourse is fundamentally in error in this regard.

Some who accept moral realism, such as Russ Shafer-Landau (1994) and Paul Bloomfield (2001), reject absolutism, claiming that moral facts can sometimes imply more than a single correct moral course of action. However, it is not my intention to critique or defend these metaethical theses, but only to suggest that the combination of cognitivism and objectivist realism – often accompanied by absolutism – is a common, and perhaps intuitively appealing, interpretation of the apparent objectivist moral language that forms a core part of our everyday moral experience.

3.1.1: Moral disagreements

The existence of moral diversity presents a potentially serious challenge to the thesis of moral objectivism. If there exist objective facts that can settle moral questions, then we might expect greater convergence of moral views and a subsequent diminishing of moral diversity than we observe in the world today. John Mackie was particularly impressed by the widespread and persistent moral diversity in the world, and he offered the argument from disagreement[4] as one of his two challenges against the truth of moral realism. The crux of Mackie's argument from disagreement is that moral disagreement might be better explained in terms of "people's adherence to and participation in different ways of life" rather than the "hypothesis that they express perceptions, most of them seriously inadequate and badly distorted, of objective values" (op. cit.). Thus Mackie suggested that moral disagreement might actually be better understood as being more like disagreement about food than disagreement about the orbits of the planets; morality does not concern the appreciation of objective facts about right and wrong that steer our moral judgements, but rather reflects the contingencies of our particular ways of living. Other philosophers have made somewhat similar claims, such as moral relativists like Gilbert Harman (Harman & Thompson, 1996) and David Wong (D. B. Wong, 2006; D. Wong, 1991). I will have more to say about the latter views in section 3.2.

3.1.2: Fundamental moral disagreement

However not all disagreement is necessarily indicative that there is no fact of the matter. As mentioned above, there are manifold disagreements in the natural sciences, for example, yet this (alone) does not readily lead us to presume there is no fact of the matter that can settle such disagreements. Disagreements in discourses that are generally agreed to be realist can potentially occur as a result of ignorance on behalf of one or more interlocutors of the pertinent facts, or it can be due to bias, or as a result of insufficient

[4] Following David Brink (1984), I will refer to the "argument from relativity" by the more accurate moniker of "the argument from disagreement." Brink points out that calling it the "argument from relativity" begs the question, as the argument starts with disagreement and arrives at the explanation of relativity rather than realism. Calling it the "argument from disagreement" does not presume the conclusion.

cognitive or reasoning capacity to determine the correct conclusions from the available evidence. But what if we account for these potentially confounding factors? What if the two agents were not mere ignorant, biased and cognitively limited mortals such as we but were perfectly rational, without bias and had full access to the pertinent facts? They would, no doubt, find themselves in complete agreement over a great many matters. But would they find themselves in perfect agreement over all their moral attitudes and beliefs? If so, then this robs the argument from disagreement of much of its persuasiveness.

As such, it is not the existence of *apparent* moral disagreement that ought to lead us to question moral realism but the existence of *fundamental* moral disagreement – disagreement that fails to dissolve between two ideally situated agents. As such, fundamental moral disagreement is the focus of much attention in the metaethical literature.

Russ Shafer-Landau (1994) gives a particularly strong formulation of the argument from disagreement aimed at focusing on fundamental, or "intractable", moral disagreement:

> 1. If there are objective moral facts, then there can be no intractable moral disagreement among ideal moral judges.
> 2. There can be such disagreement.
> 3. Therefore there are no objective moral facts.

I will call the first premise the "convergence" premise and the second the "fundamental moral disagreement" (FMD) premise. The convergence premise is often couched as a comparison between ethics and the natural sciences, as discussed in section 2.1. Most people would likely agree that the discovery of objective facts by the sciences has driven a widespread convergence in belief about how the natural world is, and that is precisely what convergence driven by objectivism and realism ought to look like. The FMD premise states that the expected convergence in belief does not exist when it comes to moral matters, as demonstrated by the existence of moral diversity and disagreement in the world. If both premises are true, then it appears that the conclusion must also be true, and objectivist moral realism is false.

There have been two common responses from those defending moral objectivism and realism against such an argument from disagreement. The first approach is to accept the convergence premise but reject the second premise concerning the existence of fundamental moral disagreement. Doris and Plakias (2008) helpfully label advocates of this view "convergentists". The second approach is to accept the second premise, but reject the first, claiming that fundamental moral disagreement does not necessarily imply a wholesale convergence of belief – earning them the moniker of "divergentists" (ibid.). Effectively, moral realists can either dismiss diversity or embrace it, although both approaches have their problems.

3.1.3: Convergentism

Convergentists, such as David Brink and Michael Smith, are committed to the notion that morality is grounded in objective moral facts, recourse to which can potentially resolve *any* moral dispute. Convergentists effectively dismiss moral diversity as being an artefact of error, thus merely a hurdle to overcome en route to the correct answers to moral questions.

While convergentists are committed to the existence of objective moral facts, they do not necessarily agree on what those moral facts look like. Brink, for example, sees these facts as being natural facts, whereas Smith sees these facts as being *a priori* moral truths accessible under conditions of "full rationality" (Brink, 1984). However, they generally agree that facts of some flavour exist that can suitably dissolve moral disagreements.

Convergentists notably concede that the existence of truly intractable moral disagreement would cast doubt on the existence of such objective moral facts. However, just because people disagree in polite (or not so polite) conversation, this alone does not imply that their disagreements are *truly* intractable and that there is no fact of the matter that could potentially settle the dispute. As mentioned above, it may be that one or more of the interlocutors is in error about some salient (non-moral[5]) fact, or that they are talking across each other, or that they are both wrong. It is important to ask whether the moral disagreement would persist even once these possible extenuating circumstances are accounted for. Thus a common strategy employed by convergentists is to appeal to such extenuating circumstances and claim that all observed disagreement is merely *apparent*.

One example is the claim that many apparent moral disagreements might be subject to what Doris and Plakias (op. cit.) call "defusing explanations", examples of which they list as being disagreement about relevant non-moral facts, partiality or bias, irrationality or cognitive impairment, or disagreement over background theory. A defusing explanation might suggest that to the untrained eye the Greeks and the Callatiae appear to be in a state of moral disagreement, but they were *actually* in agreement over the relevant moral point that one's parents ought to be treated reverentially when they die. They only disagreed

[5] It would beg the question to suggest the error involved concerns some moral fact.

over the manner by which that reverence was displayed. Or perhaps they are both in error, and another practice is the morally correct one.

A similar example of a defusing explanation in action is offered by Walter Sinnott-Armstrong (2006). He suggests two cultures could employ norms with which we might disagree, such as Vikings and Eskimos permitting the killing of their parents with the intention of either aiding them in the afterlife or aiding their communities when under extreme resource pressure, respectively. In contemporary liberal societies we would consider these acts to be morally wrong. However, that apparent disagreement might just be a consequence of our different background beliefs about what benefits our parents and community, and our different circumstances in terms of resources. As such, the moral disagreement might be "defused" given these different beliefs and circumstances. In the case of the Eskimos, there might be agreement that it is morally good to aid one's parents and community, and only kill one's parents in very extreme circumstances. Yet only Eskimos regularly experience the kind of extreme circumstances such that they have a moral norm permitting the killing of one's parents. In the Viking case, there appears to be a genuine disagreement over non-moral facts, namely on the existence of an afterlife, but there might be moral agreement that it is morally good to aid one's parents (Fraser & Hauser, 2010). As such, in order for it to qualify as fundamental moral disagreement, the disagreement must persist in the face of agreement over non-moral facts, without bias or cognitive impairment and in like circumstances.

If such defusing explanations were found to apply to all seemingly intractable moral disagreements, then it might be that the examples of cultural and moral diversity painted by Herodotus, Westermarck and many other ethnographers and anthropologists would fail to qualify as fundamental moral disagreement. Specifically, if the individuals involved were debating under "ideal" conditions, implying they were perfectly rational and unimpeded by cognitive deficits or irrational biases, and had full access to (and thus presumably agreement upon) the relevant facts, then they would find themselves in perfect agreement over moral matters. Such agreement, if it were possible in principle, would be suggestive that all observed moral disagreement is illusory. If, however, such individuals were still at loggerheads over some moral issues, then that would suggest that fundamental moral disagreement does exist.

So, does fundamental moral disagreement really exist? There is, sadly, no conclusive answer, not least because the answer leans heavily on the vagaries of empirical evidence. There can be no doubt that there exist countless volumes of anthropological studies that document the various customs and proclivities of cultures throughout the world and through history, and from these texts can be extracted examples of apparent moral disagreement. Edward Westermarck was so impressed by the pervasive variation in these beliefs and attitudes that he became a spirited advocate of a form of moral relativism on the grounds that such disagreement could not *possibly* be explained away in terms of ignorance or bias (Westermarck, 1906). It is less clear, however, whether the examples he cites should be counted as potentially *fundamental* moral disagreement. Like many anthropological texts, Westermarck documents the stated norms of various cultures, but it is not always clear as to whether these anthropological accounts can be used to extract genuine attitudinal differences over moral principles (Moody-Adams, 1997).

More recent examinations of the empirical record have attempted to hone in on specific examples of potential disagreement over characteristically moral issues. Doris and Plakias (2008), for example, reference the seminal study conducted by Nisbett and Cohen (1996) on the culture of honour in the southern United States and how that differs to the attitudes on honour held by many northerners, and find that this could be an instance of fundamental moral disagreement. Brian Leiter (2008) as well as Fraser and Hauser (2010), however, echo Moody-Adam's concern over whether Nisbett and Cohen's research on *behavioural* differences is of a kind that can enable us to extrapolate the appropriate *attitudinal* differences between the northerners and southerners in question, thus questioning whether this is a suitable case of fundamental moral disagreement.

Abarbanell and Hauser (2010), followed by Fraser and Hauser (op. cit.), offer another example of possible fundamental moral disagreement, this time of the difference between Westerners and Maya in distinguishing the moral significance of actions versus omissions, which they suggest is a case of fundamental moral disagreement. Yet, even they admit that the case is far from being conclusive, serving only to shift the burden of proof on to the realist.

However, convergentists on the whole have remained unpersuaded by the empirical record concerning moral diversity and disagreement. For convergentists can always appeal to so-called "ideal conditions" as the ultimate "defuser" of moral disagreement. Yet without access to the ideal conditions it is difficult, if not impossible, to precisely (or even approximately) know what would and would not be true from such a rarefied perspective. As Paul Bloomfield (2008)– a realist himself – points out, asking us to imagine infallible human beings "is an oxymoron at best and a contradiction in terms at worst". He goes on to ponder

> What possible bearing any empirical evidence might have on a discussion of fundamental disagreement that involves beings capable of being so much more well-informed and rational than any actual member of *Homo sapiens* could ever possibly be. We have no empirical data, for example, nor will we ever get it, about whether or not such perfectly rational and fully informed beings would have "fundamental moral disagreement."

There is also disagreement over whether these so-called "ideal conditions" are those of perfect knowledge or of perfect rationality, or both. It is also unclear whether perfect knowledge includes knowledge of the pertinent moral facts. If so, that would presumably beg the question. Does perfect rationality mean access to all the relevant facts or only mean some kind of ideal process of reasoning given limited information, as Shafer-Landau (2003) argues? Without access to these ideal conditions ourselves, or even agreement over what constitutes these ideal conditions, the disagreement over this kind of disagreement may persist indefinitely.

Furthermore, the elusive nature of the ideal conditions, coupled to the prospect that moral agreement might require such conditions, raises unsettling questions about the epistemic accessibility of moral facts. If it turns out that moral agreement rests on unattainable states, this potentially makes ethics an unpalatably impractical pursuit. We might be content with the idea that some facts about the state of the natural world may forever be beyond our grasp, but the pressing practical concern of ethics makes such a condition on ethical knowledge unacceptable for most. And if it turned out that certain key moral facts were forever beyond our ken, then this would have radical implications for the practice of moral enquiry, perhaps *encouraging* tolerance of diversity – perhaps even promoting a pluralistic moral viewpoint.

It seems as though the existence of fundamental moral disagreement poses a problem for convergentist realism. It might be a common crutch in terms of philosophical argumentation, but I would suggest the overwhelming evidence of moral disagreement in the world puts the burden of proof on the convergentists to demonstrate that such disagreement is only illusory rather than being some manifestation of a deep and abiding deficiency in human moral thinking. And I find the arguments to date in favour of fundamental moral disagreement unconvincing.

Convergentism is also a somewhat unsatisfying thesis in that it unapologetically seeks not to *explain* the existence and patterns of moral diversity in the world, but rather to *explain them away*. If convergentism is true, and fundamental moral disagreement does not exist, then the diversity of moral attitudes and norms in the world is merely an artefact of ignorance, bias or cognitive deficit. Such a thesis seems to me to be overly flippant, dismissing as it does the vast range of moral views in the world and the possibility there might be something interesting about *why* particular cultures or individuals have adopted the moral outlooks they have. Is it an artefact of ignorance that individuals living in harsh environments often have norms concerning the appropriate disposal of family members who are a burden on other members of the community, while those living in affluent temperate societies do not? Is it an artefact of ignorance that individuals living in times of conflict and war often have norms that encourage loyalty and stronger punishment for those who are disloyal while those that live in more peaceful times often have norms that are more tolerant of diversity and dissent? These appear to me to be rich and interesting questions that deserve an explanation rather than to be explained away. For this reason, and those mentioned above, I find the convergentist approach to be an unsatisfying lunge for certainty at too high a cost.

3.1.4: Divergentism

The second response from moral realists to the argument from disagreement as it is outlined in 3.1.2 is to acknowledge that some disagreements may indeed be genuinely intractable, but to deny that this undermines moral objectivity or realism. Thus they reject the convergentist premise of Shafer-Landau's argument and accept the FMD premise.

Robert Boyd (1988), for example, offers what he calls a "homeostatic consequentialist" theory of good, whereby "good" is defined as a cluster of features that tend to "satisfy important human needs" (which are left unelaborated). The virtue of such a definition, according to Boyd, is that it can serve as a natural definition without the strictures of satisfying a fixed set of necessary and sufficient conditions. Rather, a "homeostatic property-cluster definition" takes a family of properties that tend to cluster together in a self-supporting manner. This is reminiscent of prototype theory in semantics (Osherson & Smith, 1981), whereby a concept is not defined in terms of necessary and sufficient conditions, but by a cluster of features that tend to appear together, triggering recognition when present in sufficient quantity. Boyd uses the concept of species from biology as an example of an approach that employs homeostatic property clusters: individual organisms of the same species feature significant morphological, physiological and behavioural similarity, while allowing for borderline cases or populations that are intermediate between two other species. This approach to defining the good actually predicts that there will be borderline and other “hard” cases in ethics, where various features of the good conflict without calling into question the reality of goodness itself.

However, such a sophisticated homeostatic consequentialist account diverges from what most users of moral language probably think they are talking about. To the extent that a metaethical theory is supposed to represent what it is that people are doing when they employ moral language, this could be considered a setback for homeostatic consequentialism. Furthermore, like species, there are probably discoverable patterns and systems that can at least allow people to agree to disagree given a mutually understood vagueness that exists between borderline cases. Yet many moral disagreements do not appear to concern borderline cases: it either is or is not permissible to kill people as a form of punishment. It would take some work to demonstrate how this form of moral disagreement is due to it concerning some borderline issue.

Russ Shafer-Landau (2003), takes a different tack, and argues that even in ideal conditions agents might continue to disagree, but that this does not necessarily undermine objectivism. This is because the views of even idealised agents don’t *constitute* the truth of the matter, and even idealised agents may be in error or ignorant of some pertinent facts. This stems from Shafer-Landau’s notion that perfect rationality is procedural: “being rational or exercising one’s rationality essentially involved a series of operations over one’s existing commitments”. Even furnished with the relevant facts, idealised agents might disagree unless they were also stripped of any individual circumstances that might introduce biases. Yet stripping away individual circumstances might be anathema to moral disagreement; after all, it is not too difficult to believe that if we were all similarly constituted and in similar circumstances, with similar non-overlapping interests, then we might all be of a like mind. Yet the nature of moral disagreement is that individuals are in different circumstances with conflicting interests and subjects of interest. Explaining that aspect of our moral experience away seems a touch expedient.

Even so, Shafer-Landau does not consider fundamental moral disagreement, even under ideal circumstances, to be a showstopper for moral realism. He also posits that there might be “truth value indeterminacy”, which suggests that some moral judgements are simply neither entirely true nor entirely false (Shafer-Landau, 1994). Or there could be “comparison indeterminacy”, such as cases of genuine moral dilemmas, where competing outcomes are morally equivalent with no way to choose between them. David Brink (1984) has also offered a similar view, suggesting that some moral facts might be contingent, and as such, some moral disputes have “no uniquely correct answers.” He concedes that “moral ties are possible, and considerations, each of which is objectively valuable, may be incommensurable.” These approaches allow for the existence of objective moral facts, but these facts are not necessarily always in accord, and can suffer from vagueness, indeterminacy, equivalence or contingency, with moral disagreement being symptomatic of these phenomena.

This raises the possibility of another approach to dissolving moral disagreement, at least in a sense: both interlocutors could just *agree to disagree*. Presumably, in ideal circumstances, they would at least be able to acknowledge the pertinent indeterminacy or equivalence, and concede that their position is not the only one on this moral issue. A genuine “moral tie” might be recognised as such, thus representing an agreement of sorts. However, such fundamental agreement-to-disagree seems to fail to capture one of the signature characteristics of moral disagreement, which is a perceived incommensurability between what two individuals believe to be the right answer, with no apparent way of finding agreement, or even agreeing to disagree. Furthermore, given the difficulties raised in the previous section in knowing quite what ideal agents might believe, it is difficult to know whether idealised agents would agree, disagree or agree to disagree.

Ultimately I find this kind of divergentism not substantially dissimilar from convergentism in that both call for us to imagine what we people in ideal conditions might agree or disagree upon. Until objectivists give us a better idea of what such conditions look like, and what people in those conditions might agree upon, then we ought to remain sceptical about any argument that depends on these rarefied conditions.

3.1.5: Bloomfield’s divergentism

An alternate divergentist approach is taken by Paul Bloomfield (2001), who draws an analogy between realism about morality with realism about health. It is generally agreed that there are objective facts about being alive, dead and all the various states in between. This constitutes a realism of sorts. Yet there can still be disagreement about what constitutes a healthy course of action compared to one that hampers health. Genuine and intractable disagreements can then emerge about things such as what amount of body fat is “healthy,” but such disagreements do not call into question the existence of the facts underlying weight, metabolism, mortality and so on. Bloomfield argues that if two individuals disagree over a functionally equivalent point – meaning there’s no significant difference between them in terms of measurable impact on fitness or mortality – their disagreement becomes one of convention rather than representing intractable moral disagreement. In this way, Bloomfield accounts for much of the moral disagreement in the world:

> Concepts of *health* may vary from one culture to another and be functionally equivalent. If they are not functionally equivalent, what differentiates them is not merely convention: these would be differences that express themselves as variations in fitness and mortality rates in those cultures. Merely conventional differences in health have no functional effect; if there are functional differences, then we do not want to say that what causes them is merely conventional.

That said, if their disagreement persists after conventions have been dissolved, such as if two cultures disagreed on whether mortality rates were an appropriate metric for health, then this disagreement would appear to be more fundamental than Bloomfield's account of conventional disagreement. Doris and Plakias (2008) also point out that there might well exist cases of genuine fundamental disagreement within health, particularly in cases of mental health. They reference the debates that are still ongoing particularly in mental health concerning diagnosis, let alone the debates concerning treatment.

Richard Joyce (2003) offers another criticism of Bloomfield's analogy between health and morality. Discussions in morality and health both have a prescriptive element, but the strength of this prescriptive aspect is crucially different. Joyce sees morality as necessarily possessing a practical clout that holds "irrespective of one's desires", as demonstrated by cases where someone might be fully informed about the facts of how an act of torture will advance their ends, yet we might argue that they are morally wrong to engage in torture despite their true beliefs. Joyce questions whether Bloomfield's account of health includes such a crucial appeal to practical clout that makes the analogy with morality salient. This clout might conceivably be provided if an individual *desires* health, but there do not appear to be any facts *about* health that bind an individual to desire health. Thus, if an individual opts out of valuing health, all the prescriptive claims about how one ought to behave in order to be healthy fail to apply in a way that opting out of morality does not make moral norms fail to apply. Thus, if morality is analogous to health, then this relegates morality to a kind of system of hypothetical imperatives – to put it in Kantian or Footian parlance – that only bind to the extent that an individual desires the ends provided by morality.

While there are good reasons to think Bloomfield's health analogy fails to bail out moral realism, in its very failings it might be rescued as a defence of a functionalist *anti-realist* account of morality, similar to the one I will offer in this thesis.

Divergentism is a more nuanced position than convergentism, affording more sophisticated accounts of moral realism to account for the existence of fundamental moral disagreement. Yet even divergentists would likely agree that fundamental moral disagreement can only be so prevalent before moral realism is called into question. Shafer-Landau, Brink and Bloomfield all agree that there ought to be a considerable amount of fundamental moral *agreement*, even allowing for indeterminacy, contingency and functionally equivalent conventions. This raises again the question mentioned in 3.1.2 of the existence and extent of fundamental moral disagreement – a question that is devilishly difficult to answer decisively one way or the other by recourse to the empirical record.

3.2: Moral subjectivism

One coarse way to carve up the metaethical spectrum is in terms of whether someone is more impressed by the apparent objectivism in our moral language or by the existence of moral diversity in the world. Above I dealt with some theories from those impressed by the apparent objectivity of morality, and in this section I will look at three theses from individuals more impressed by the existence of moral diversity and disagreement. The first, in opposition to moral objectivism, is moral *subjectivism*. This states that morality is not constituted by objective facts but by our subjective attitudes on some level. A particularly bold (or bald, depending on how you look at it) form of subjectivism would state that moral statements are made true purely by recourse to an individual's subjective attitudes. However, such a position would seem rather untenable in practice, as it would open the door to anything an individual happens to prefer to be the morally right thing to do. The consequence could conceivably be an anything goes free-for-all that would appear to be anathema to the presumed action guiding nature of morality.

Instead, subjectivism is often couched in other terms, such as the non-cognitivist theories, like emotivism. A. J. Ayer (1936), for example, argued that moral utterances are not the statements of fact that they appear to be, but rather are expressions of attitude. Thus, when someone says "torturing innocents is wrong", they really mean that they hold a feeling of disapproval towards the action of torturing of innocents, and they make an assertive utterance to that effect, compelling others to feel the same. Such a position has the virtue of being able to account for the apparent motivating or "action guiding" force of moral utterances by appeal to the emotional compulsion that constitutes moral judgement. However, expressions of emotion or attitude are not the kinds of things that can be found to be true or false. This means moral statements are also not the kind of thing that can be found true or false, except in the descriptive sense that it is true or false that a particular individual holds a certain moral attitude.

Such non-cognitivist and emotivist theories were quite popular in the mid-20th century, largely in reaction to the apparent intractability of accounting for the evaluative nature of the moral facts or properties that were seen to underlie realist interpretations of moral utterances. Yet, while non-cognitivism stressed that moral utterances contain some evaluative component that stirred us into action, they suffered from a number of other flaws. One such is that expressive statements might work in a fairly simple form, such as "torture is wrong" or "you ought not torture," but it is rather more difficult to make them work in other ways in which we use them. Chiefly, expressive statements without factual (i.e. propositional) content are difficult to reconcile with conditional statements, such as "if lying causes harm, then it is wrong to lie," which are not uncommon in moral discourse – an issue referred to as the Geach-Frege problem after P. T. Geach brought it to light (Geach, 1965).

As mentioned above, John Mackie rejected the idea that moral utterances could best be understood as being *purely* expressions of emotion or preference. Doing so ignores the apparently objectivist nature of much moral discourse, including the rich tradition of moral argumentation, requiring something of a radical revision of our moral language. Emotivism further undermines moral argumentation by removing the recourse to reason when it comes to disputes over values. Ayer asserted that when one engages in moral debate they are only debating the pertinent non-moral facts about an issue or a particular case. He denied it was possible to engage in rational debate over which values were the right ones, which is why disagreements over value often devolve into "mere abuse" (op. cit.). And while emotive expressions might prove persuasive, it seems wrong to make the big step to declare that the most persuasive emotive expressions deserve to be called the right ones.

The second thesis, in opposition to moral realism, is moral *anti-realism*. Some, such as John Mackie (1977), Simon Blackburn (1993), Richard Joyce (2001) and Joshua Greene (2002), have rejected the realist interpretation of moral discourse by questioning the existence of objective moral facts. They argue that moral discourse is objectivist by nature, which would imply an underlying realism, but that the facts the discourse purports to reference simply do not exist. This constitutes a broad "error-theory" of moral discourse, similar to how an atheist might consider religious discourse referencing supernatural beings to be fundamentally in error.

In claiming that moral facts do not exist, and as a consequence that there is no fact of the matter about whether a certain act is right or wrong, error-theories are sometimes thought to lead to a form of moral nihilism, with subsequently disconcerting implications for those of us (perhaps all of us?) who are reluctant to dismiss morality as an illusion or error. Joyce's response to this concern is to advocate a form of moral fictionalism, whereby we acknowledge that objectivist moral discourse is in error, but we continue using it anyway because it is useful in steering our behaviour in (what we used to call) "moral" ways. Joyce is particularly interested in the binding prescriptivity – the "practical clout" – that morality has, and he is concerned that stripping moral discourse of its presumed factual basis might undermine morality's practical clout. Without the overridingness of moral norms, we might succumb to other prudential or selfish concerns and cease behaving morally. Thus we continue to employ moral realist language to help compel not only others but also ourselves to behave "morally." However, fictionalism complicates the debate about which moral norms we should adopt, and on what basis we should adopt and defend them, given there is no fact of the matter that makes one position true and another false.

A somewhat similar response to the implications of moral anti-realism was offered by Simon Blackburn several years prior to Joyce's account. Blackburn's response was to revise the way we interpret and use moral terms. He builds an elaborate semantics that incorporates expressivist terms that can be used in place of objectivist terms, so "yay!" could replace "is good" and "boo!" could replace "is wrong," for example (Blackburn, 1984). One benefit of this system is it can respond to the Geach-Frege problem mentioned above by enabling expressivist terms to accommodate conditionals. Blackburn asserts that such a fully elaborated system could plausibly replace our objectivist discourse without making any serious changes to the way we actually engage in moral conversation. Blackburn dubs his thesis "quasi-realism," in that it happily preserves many of the tropes of realist discourse without the questionable metaphysical grounding.

However, as a fellow anti-realist, Joshua Greene points out, even if everybody engaging in moral discourse makes the fictionalist or quasi-realist "turn" – and acknowledges with a wink and a nod that they're *really* talking about expressions of pro or con attitudes or some other idea rather than referencing moral facts – the problems of moral realism will likely persist. These problems occur when two moral systems rub up against each other, and individuals from both sides have difficulty reconciling their views because they see their own system as being made true by recourse to facts, rather than them *really* being expressions of their pro or con attitudes. Conflict – often bloody – ensues.

Greene shares Blackburn's view that people tend to project their moral attitudes on the world, with a corresponding belief that their normative system is the right (i.e. true) one, and backs it up with some evidence from recent moral psychology studies. As a result, the "normal" brand of moral discourse that arises from this psychology is unrecoverably riddled with realist assumptions. As such, Greene suggests that continuing with business-as-usual realist discourse will likely result in similar conflicts occurring. Greene argues that it is not the preservation of realist moral discourse that is the objective of the anti-realist, but a revision of moral discourse to entirely rid it of erroneous realist terms.

Green's response to the spectre of moral nihilism, and the inadequacies of fictionalism and quasi-realism, is to recommend that we thoroughly revise moral discourse along constructivist lines. Instead of us talking about morality as being *discovered*, we acknowledge that it is *created*. Thus, instead of talking about morality as being justified by objective facts, we talk of moral norms as being justified by recourse to some agreed upon constructed system of norms. Such systems already exist, such as those proposed by John Rawls (1972) or David Gauthier (1986). Earlier still, Plato gives a nod to constructivism as being one way to think about morality (if not the way he preferred to think about morality) in the words of Glaucon in *The Republic*.

If realism really is untenable, then Greene's call for a revisionist constructivism seems warranted, although one could ask what would inform such a constructionism. What is the goal of morality? What is its function? What norms best satisfy that function? In looking at morality through the lens of evolution in the following chapters, I will be considering many of these questions. Once they have been addressed, it is possible the answers could help provide some foundations for a revisionist constructionism, a point I will revisit in the final chapter.

3.3: Moral relativism

A third thesis inspired by the existence of moral diversity and disagreement, and one that sits in opposition to moral absolutism, is moral *relativism*. Relativism itself can be cashed out in a number of ways, with varying degrees of sophistication or consistency. In essence, relativism describes the view that moral norms are made right or wrong by recourse to some particular non-universal framework; Gilbert Harman likens his brand of relativism about right and wrong to Einstein's relativism about mass and time (Harman, 1991). As such, according to moral relativism, norms are indexical, but a central question becomes: indexed to *what*?

Relativism is more often employed as a pejorative than offered as a genuine moral standpoint, largely because it is often conceived in an overly flippant way or is summarily misunderstood. Many folk references to relativism – often cited by proponents of one form of another of objectivism, such as theism – paint it as an anything goes free-for-all. In such a bald form of relativism, moral norms are indexed to each individual's subjective preferences and attitudes. This becomes little more than a form of ethical egoism, whereby one's actions are not only influenced by their preferences but are justified by them. However, such anything-goes relativism seems to strip morality of its moralising component – the part that is directed towards judging and steering the behaviour of others, or Joyce's "practical clout" – to the extent that moral justification and judgement become almost meaningless. I have yet to see any serious proposals by philosophers for any such kind of moral relativism.

Another rendering of moral relativism indexes moral norms to a particular cultural framework. This brand of relativism seems particularly popular in postmodern texts, which appear to be motivated by an imperative to avoid the evils of colonial imperialism, whereby one culture imposes its moral system on another under the pretext that it is the "right" one. This is closely tied to a social constructivist theory of knowledge (such as that proposed by Berger & Luckmann, 1966) which takes not only right and wrong but facts themselves to be socially constructed. Coupled with that is a deep suspicion of those in power who can, and do, shape the construction of facts and values to serve their own ends, institutions such as cultural elites, the media, big business, big religion, etc. Combine these theses and it becomes clear why a form of cultural relativism might emerge. Facts and values become suspect, as they are manipulated by the elites in order to preserve their status. As there is no objective right or wrong from this perspective, then one culture cannot claim to have a privileged moral system compared to another culture, and thus has no right to impose its moral system on the other. However, such a moral relativism is heavily contingent on a particularly radical view of the social construction of not only knowledge but reality itself. It is also largely ignored in the metaethical literature. I will not confront such a relativism in detail in this thesis besides give it a cursory nod before moving on to more substantiative moral relativist theories.

Edward Westermark, for example, famously articulated what he saw as a profound descriptive moral relativism in his two volume *The Origin and Development of the Moral*

Ideas (1906), and developed it into a strong normative relativism in his follow-up book *Ethical Relativity* (1932). Two more recent vocal proponents of moral relativism are Gilbert Harman and David Wong. Both have broadly similar renditions of relativism, which differ in detail, but are both committed to the notion that there is no one single correct or true moral system that applies to all people at all times. Harman proposes that

> Moral right and wrong (good and bad, justice and injustice, virtue and vice, etc.) are always relative to a choice of moral framework. What is morally right in relation to one moral framework can be morally wrong in relation to a different moral framework. And no moral framework is objectively privileged as the one true morality. (Harman & Thompson, 1996)

Harman goes on to argue for a form of constructivist moral conventionalism, whereby morality is created by humans for the purposes of facilitating social living, and that "right" and "wrong" should be considered relative to the one particular constructed moral system – not unlike Greene's proposed constructivism mentioned above.

David Wong offers a similar account:

> Human beings have needs to resolve internal conflicts between requirements and to resolve interpersonal conflicts of interest. Morality is a social creation that evolved in response to these needs. There are constraints on what a morality could be like and still serve those needs. These constraints are derived from the physical environment, from human nature, and from standards of rationality, but they are not enough to eliminate all but one morality as meeting those needs. Moral relativity is an indication of the plasticity of human nature, of the power of ways of life to determine what constitutes a satisfactory resolution of the conflicts morality is intended to resolve. (D. Wong, 1984)

For Wong's "pluralistic relativism" (D. B. Wong, 2006), morality is relative – or indexed – to the environmental conditions, including the "availability of human and material resources" (ibid.), which defines the nature of the problems that human beings want to solve in order to satisfy their needs. However, crucially, there is no single solution that works in all environments. This puts relativism at odds with the idea that there are objective facts that make certain actions right or wrong for all people at all times. However, it does not necessarily put relativism at odds with a form of moral realism that states that the circumstances, broadly interpreted, are highly relevant to moral judgements, and these circumstances include things like the environmental conditions. Likewise, Wong believes there can be some tenets that all moral systems will likely endorse if they are to advance the ends of promoting social living, cooperation and human flourishing. Encouraging random violence, or compulsory lying, for example, are likely to be thoroughly poor norms in any environment.

One strength of Wong's pluralistic relativism is that it does not fall foul of claims it endorses an anything-goes brand of moral relativism. He places boundaries on what could constitute an acceptable moral code, eliminating things like bald egoism or Nazi moralities. It does, however, leave some issues to be resolved, such as how an individual with overlapping cultural or group identities ought to reconcile their moral systems if they conflict, or how an individual with a moral code adopted from one environment ought to adjust their moral views when moving into a different environment, or how a moral system itself ought to adjust if its environment changes.

Still, I am sympathetic (to a point) with Wong's pluralistic relativism, and there are some parallels with the story of moral ecology that I will elaborate throughout this thesis. Particularly relevant is the notion that "right" and "wrong" are, in a sense, contingent on environmental circumstances, as I will discuss in later chapters.

3.4: Subjectivity and inescapability

Having characterised a number of views that qualify as broadly subjectivist, it is worth stressing that there are many people who are concerned that any brand of subjectivism is doomed to fail as a moral system *qua* a moral system. I mentioned above that one of the core – perhaps defining – features of morality is that moral norms unconditionally bind an individual irrespective of their attitudes, desires or interests. This is captured in the idea of a moral norm as a categorical imperative, as elaborated by Kant, a feature that even moral sceptics such as John Mackie and Richard Joyce believe is central to any coherent interpretation of moral language.

Moral realism caters for this feature of morality by placing the grounds of justification outside the attitudes, interests or desires of the agent and rooting them in objective features of the world. In the same way that someone cannot desire that the orbits of the planets are circular, and use that attitude to justify their assertion that the orbits *are* circular, someone cannot desire that lying or torture are permissible and then use that attitude to justify lying or torturing. Rather, the realist asserts that it corresponds with our usage of moral language that there must be some objective fact of the matter that provides a firm foundation for whether a particular action is permissible or otherwise.

Subjectivism thus is often criticised as undermining the binding authority of morality, such that morality loses its potency in guiding our behaviour and the behaviour of others. If morality weakens to “mere” convention, or becomes predicated on subjective preferences, then what is to provide the motivating force to adhere to its strictures? What is to prevent someone from expressing a desire to pursue their own interests unhindered by the bounds of morality, thus precipitating a slide into moral anarchy or nihilism? A similar concern is expressed in the popular quote from Dostoyevsky’s *The Brothers Karamazov* that “without God, everything is permitted” – which could be translated in this context as saying “without moral objectivism, everything is permitted”. To the extent that morality is required to possess this binding prescriptive character, this is a serious charge that subjectivism must overcome. I will return to the notion of binding authority in chapter 17 when I look at the implications of moral ecology on metaethics.

3.5: Best explanation

So far in this chapter I have provided a definition of the phenomenon of moral disagreement and discussed its implications on two metaethical approaches: realism and subjectivism. However, there is a broader point to make about the importance of moral diversity beyond its implications for moral realism.

I would suggest that moral diversity is actually a far more interesting phenomenon than the discussion about fundamental moral disagreement alone might imply. Because of the emphasis on fundamental moral disagreement within metaethics, discussions of moral diversity in the literature often focus on a peculiarly narrow range of issues, centring around the truth or otherwise of a particular brand of moral realism. On the one hand we tend to have realists who typically seek to explain moral diversity away as being some artefact of ignorance or cognitive shortcoming. On the other hand we have subjectivists and anti-realists who claim that fundamental moral disagreement does exist and it puts paid to the possibility that there exist objective moral facts. Both scour the empirical record for cases of intractable moral disagreement and attempt to reckon whether they can be defused or might dissolve in an ethereal discussion between idealised moral agents.

However, this is just one way to look at the phenomenon of moral diversity and its significance on morality. In fact, I would suggest that the debate about fundamental moral disagreement misses much of what is interesting about moral diversity. This is because it focuses on a particularly rarefied version of moral disagreement rather than looking at moral diversity as being a phenomenon worthy of exploration and explanation in and of itself. After all, even if there do exist moral facts, and moral diversity is merely an artefact of error, the question still remains as to why such errors exist and how it is that people come to the erroneous moral beliefs that they do. Even if there are verifiably correct answers to moral questions, as Michael Smith suggests there are, and even if we happen to have stumbled upon them, the task remains to convince others that these are, indeed, the correct answers. As such, it might pay to understand how it is that people come to the moral beliefs that they do, and how they can be persuaded otherwise.

Perhaps an even more interesting aspect of the phenomenon of moral diversity is looking at whether there exist identifiable patterns in the diversity among moral systems. Random variation is often rather uninteresting. Structured variation, on the other hand, is often a sign of some interesting underlying regularity. Variation in biology, for example, is often not just the product of random noise, or indicative of an organism failing to track some fixed objective standard, but is a response to the contingencies of the environment in which that species evolved. Variation in size, shape and feeding habits in Galapagos finches, for example, was an enticing hint that sparked Darwin to the idea that species evolved in response to the relative abundance of various food sources in the environment in which they live (or in which they have evolved).

In parallel, perhaps moral diversity and disagreement are similarly indicative of variation in what Mackie calls "ways of life", such that particular moral attitudes are more likely to emerge given certain patterns of living or in certain environmental conditions. And perhaps individuals who adopt those ways of life are quite rational – rather than being irrational or grossly in error – in holding the moral attitudes they do. Perhaps those moral attitudes help them to live successfully in their particular environment. This alone does not necessarily imply that moral diversity is not a product of ignorance or bias – after all, error might also vary in accordance with ways of life or environmental contingencies – but if a story can be told that makes sense of the connection between "ways of life" and morality, then the explanation might support the conclusion of Mackie's abductive argument that moral objectivism and realism are not the best explanation of our moral experience. I would suggest that it is in this context that moral diversity is most interesting and worthwhile of further study.

In fact, it seems rather anomalous that Mackie's argument from disagreement is often rendered in a form that is quite different from the one offered by Mackie himself, in a way that meanders off into the distraction of hunting for fundamental moral disagreement. Shafer-Landau's argument offered in 3.1.2, for example, is in deductive syllogistic form, whereas Mackie explicitly offered his in abductive form: as an argument from best explanation. Mackie sought with his argument from disagreement not to land a knockout blow against moral realism on the basis of the empirical evidence for moral diversity, but rather suggested that the realist explanation of the phenomenon of moral disagreement might not be the *best* one given what we know. Mackie was less interested in debating the possibility of the existence of fundamental moral disagreement and, in lieu of its discovery, was more interested in what the current empirical record on moral disagreement could tell us. In this case, it is a matter of weighing up the evidence for either explanation, and making a judgement over which explanation is more likely.

Gilbert Harman has made a similar point that moral diversity in and of itself does not *entail* moral relativism, "any more than Einstein claimed that differences in opinion about simultaneity by themselves entailed relativistic physics" (Harman, 1991). Likewise, observable phenomena do not tend to entail their explanations. Rather, phenomena typically have multiple possible explanations, but we tend to seek the explanation that is best able to account for the breadth of the phenomena while remaining consistent with our other best explanations for other related phenomena.

Returning to Mackie's abductive argument, the crucial point he makes is whether moral diversity is better explained by "people's adherence to and participation in different ways of life" rather than it being symptomatic of most or all people being deeply mistaken about their moral beliefs. In the absence of some conclusive story about the existence of fundamental moral disagreement – or the discovery of some idealised rational agents whom we might interrogate about the matter – it seems that Mackie's abductive argument is still the best way to frame the debate about the implications on objectivist realism of moral diversity in the world. As such, the key to the issue is to cash out what is meant by "ways of life". This thesis can be seen as an attempt to do just that. It ultimately defends the view that understanding "ways of life", particularly in terms of environmental contingencies, is crucial to understanding the nature of morality, which suggests a kind of sophisticated subjectivism.

In the remainder of this thesis, I will argue that moral diversity is not just some anomaly caused by ignorance, bias or cognitive limitations. Rather, moral diversity is a core feature of morality, it being a product of adapting behavioural codes employed to facilitate social and cooperative behaviour in a wide range of environments. To the extent that morality serves the function of facilitating social living (as I will argue it does in chapter 6), and to the extent that there is no single solution to the many problems presented by this goal, then diversity should be expected to emerge. To the extent that these things are the case, then diversity might actually be useful: some problems have multiple functionally equivalent solutions; some are trade-offs; sometimes a pluralism of behavioural strategies co-existing will promote greater levels of prosocial and cooperative behaviour than a single strategy working in isolation; sometimes the solutions innovated will be sub-optimal; and sometimes those in power will corrupt morality to serve their own interests. Moral systems, as cultural innovations, are intricately linked with the contingencies of the broader culture. Culture, in turn, is a response to the problems of living in the environment in which it exists. And our minds, having evolved over thousands of generations confronting the complex problems of social living, have been shaped to be variable and sensitive to changing and dynamic social conditions. Thus, moral diversity is not just something to be explained away; understanding its dynamics can yield insights into the nature of morality and us as moral animals. The first step in painting this picture of morality is to clarify the definition of morality itself, which will occupy the next chapter.

Chapter 4: Morality Inside-out

> Moral science is not something with a separate province. It is physical, biological and historic knowledge placed in a human context where it will illuminate and guide the activities of men.
> - John Dewey

4.0: Moral phenomena

In this chapter I will lay some important groundwork for the broad, or “outside-in,” perspective on morality that I take in this thesis, and contrast it to the prevailing narrow, or “inside-out,” view. In the next chapter I will go on to discuss the bounds of the moral domain according to the two views, and argue for the validity of expanding the moral domain to include many phenomena that are often considered to be non-moral, such as psychological forces, conventions and customs that contribute to producing prosocial and cooperative behaviour, and some of the norms that promote such behaviour. In chapter 6 I will offer a functionalist definition of morality that dovetails with the “outside-in” view, a definition that will form the foundation for my articulation of moral ecology in later chapters.

4.1: Social nature

Of *Homo sapiens’* many impressive qualities, perhaps the one most remarked upon is our unparalleled cognitive capacity. Yet an equally remarkable trait – and one that is intricately linked with the aforementioned cognitive endowment – is our spectacularly social nature. No other creature engages in such rich and complex social interaction with unrelated members of its own species. Not even the eusocial insects – ants, termites and bees – engage in social and cooperative interactions with individuals who are not close kin. And those creatures that do engage in mass social behaviour with distantly or unrelated individuals, such as pack hunting wolves, schooling fish or herding wildebeest, do so according to relatively simple calculus to advance their individual benefit – a phenomenon called “mutualism” in biology – rather than out of any genuinely magnanimous tendencies (Ridley, 1996; West, Griffin, & Gardner, 2007). In short: no other creature enjoys such a rich, layered social existence as do we.

One of the most notable features of our social proclivities is a tendency to act in a way that appears contrary to our immediate interests in order to advance the interests of other individuals. Such behaviour – often defined as "costly helping" or "altruistic" (ibid.) – was long a conundrum for evolutionary biologists, particularly when the behaviour appeared to be costly to the reproductive fitness of the individual performing the act. However, many of the evolutionary problems posed by altruistic behaviour were solved through the late-20th to early-21st century through the elaboration of mechanisms such as kin selection/inclusive fitness (Hamilton, 1963; Maynard Smith, 1964), direct reciprocity (Trivers, 1971), indirect reciprocity (Alexander, 1987), network reciprocity (Nowak & Sigmund, 2005) and multi-level selection (Wilson & Sober, 1994; M. Nowak & Highfield, 2011). The upshot is that helping another individual, even at a short term cost to oneself, can often reap a fitness benefit, either to oneself or to one's genes. If the fitness benefit outweighs the cost, then such costly altruistic behaviour can be maintained, and indeed spread, via evolution (Dawkins, 2006).

The success of these processes appears to be responsible for endowing our species with a suite of psychological mechanisms that encourage such costly helping or "altruistic" behaviour. These include emotions such as shame, empathy and guilt (Greene et al., 2001; Haidt, 2001; Huebner et al., 2009), with heuristics such as incest avoidance (Lieberman et al., 2003; Lieberman, 2008), social imitation and in-group conformity (McElreath et al., 2003), and perhaps even an innate cognitive "module" geared for navigating social interactions (Hauser, 2006). However, while we might be biologically predisposed towards at least some degree of selective psychological altruism – as defined in more detail below – our species is not immune to committing what Philip Kitcher (2011) calls "altruism failures". These are situations where our self-interest or some other confounding factor causes behaviour that hampers cooperation or harms the interests of others.

Some of our evolved tendencies to act in altruistic ways might also contribute to altruism failures. For example, kin selection – which can promote costly behaviour that advances the interest of one's close genetic relatives – can manifest in several ways. The one that receives the most attention is costly behaviour that directly or indirectly benefits kin. Another, which receives considerably less attention, is in the form of costly behaviour that directly or indirectly imposes a cost on those competing with kin. Thus the calculus behind "Hamilton's rule" that underscores inclusive fitness can also account for instances of out-group discrimination, particularly where that out-group is perceived to be composed of non-kin who threaten the interests of kin.

Our very propensity towards in-group favouritism and out-group discrimination (Haidt & Kesebir, 2010) is another tendency that appears to be the product of evolution, and which can be responsible for altruism failures, particularly as groups begin to rub up against other groups. A modicum of in-group favouritism might well have promoted cooperation within groups, not least by encouraging trust of in-group members. However, this also reinforces the need to effectively identify in-group members and discriminate them from out-group members, particularly if they are less likely to be as trustworthy as in-group members. Out-group discrimination might thus be an effective mechanism to prevent defection in cases of one-off interactions, particularly with strangers. Yet there are also benefits to expanding the cooperative circle to include neighbouring groups; a phenomenon that characterises the grand sweep of human history is the expansion of cooperation from smaller to larger populations. Barriers to this expansion – such as in-group favouritism and out-group discrimination – might thus be a source of altruism failures that can reduce cooperation overall.

4.2: Cultural nature

Our evolved psychological tendencies toward limited altruistic behaviour do not exhaust our remarkable social psychology. Another feature is our propensity to create, share and absorb information socially in the form of culture. While I will discuss culture in more detail in chapter 6, suffice to say that our capacity to share information socially has been one of the evolutionary breakthroughs that has enabled our species to range beyond the cradle of the African savannah and adapt to disparate environments in nearly every corner of the globe (R. Boyd & Richerson, 1985; Richerson & Boyd, 2005). Indeed, the power of culture to aid in our survival and reproduction appears to have placed a powerful selective pressure on our psychology to be more amenable to social transmission (Sterelny, 2012).

One of the central features of culture that is of interest to this enquiry is specifically our propensity to create, propagate and conform to behavioural rules, or “norms” (Bendor & Swistak, 2001; Buckholtz & Marois, 2012; Sripada & Stich, 2005; Sripada, 2005). Such behaviour is ubiquitous in our species, and can even be observed amongst very young children during play as they spontaneously create rules for their games, monitor conformity and sanction those who disregard the rules (Gopnik, 2009). Naturally, there is no agreed upon definition of what precisely constitutes a norm (Dubreuil & Grégoire, 2012). However, what is broadly agreed upon is that norms are rules that specify which actions are permissible or impermissible, and they tend to attract disapproval and punishment from others when breached (Sripada & Stich, 2005; ibid.). And, crucially, many norms not only concern how we ought to behave in social situations but also encourage prosocial, cooperative and altruistic behaviour.

It is the observed diversity throughout the world in these rich social and deliberative tendencies, specifically as they manifest in attitudes, behaviours and in the content of norms that appear to promote social, cooperative and altruistic behaviour, that I hope to offer an explanation in this thesis. These phenomena include: costly helping, i.e. “altruistic” behaviour; attitudes promoting such “altruistic” behaviour; the creation, adherence and punishment of norm breaches; many customs and conventions; the psychological tendencies and dispositions that contribute to our social and cooperative interactions; and the language that we employ to deliberate about such social and norm-driven interactions.

It is worth noting that, excepting the opening of this chapter, I have yet to employ the term “moral” to apply to these various phenomena. Yet I am inclined to bundle them together under the banner of “moral phenomena” rather than just calling them “pro-social” or “cooperative phenomena.” One might object at this point that such a bundle of phenomena under the banner of “morality” is overly broad, given that it includes many features that many philosophers might classify as being merely social, conventional, prudential or personal. Yet, as I will elaborate in this and the next chapter, such insistence on drawing a sharp distinction between the moral and the social, or the moral and the conventional, is an artefact of a particular perspective on morality that has been prevalent in ethics particularly over the past two centuries. This narrow, or “inside-out,” perspective has led to the moral domain shrinking to include only a limited range of behaviours and norms, namely those concerning harm and fairness. Yet it is in widening our scope, and broadening the moral domain to include a greater range of social phenomena and norms, that we can gain a deeper insight into the origins and dynamics of many of our moral proclivities, and in doing so, perhaps even resolve some of the longstanding debates that are endemic to contemporary ethics and metaethics.

4.3: Morality inside-out

There are at least two ways of attempting an analysis and explanation of moral phenomena. The first is to start with observations of our everyday moral experience and reflection upon our ordinary moral discourse. This seems a natural enough springboard for starting moral enquiry, because much of our moral experience is characterised by either introspection upon the reasons for acting in a certain way or discourse with others over the right ways to act. Much of this discourse – internalised or expressed – employs terms such as “good” and “bad,” “right” and “wrong,” “ought” and “moral,” in seemingly special ways that bear explanation. If we are to make assertions about what is “good,” then presumably we had best get straight what we mean by such terms before we begin. Some, such as A. J. Ayer, even went so far as to state,

> A strictly philosophical treatise on ethics should therefore make no ethical pronouncements. But it should, by giving analysis of ethical terms, show what is the category to which all such pronouncements belong. (Ayer, 1936)

By this, Ayer is giving priority to the study of metaethics – particularly the study of the definitions and meanings of ethical terms – rather than the study of how people ought to behave, at least for those doing philosophy in Ayer’s “strict” sense. This metaethical approach draws on the rich philosophical traditions of introspection and conceptual analysis, and brings to bear the penetrating tools of philosophy on our everyday moral language in the hope of making explicit the concepts that underlie them. Once those concepts are made clear, we presumably are better equipped to employ them correctly. Because this approach starts with our everyday moral experience and ordinary moral discourse, and then aims to work its way *out* to explain the nature of moral phenomena. I call this the “inside-out” approach to ethical enquiry.

Characteristic of the inside-out approach is reflection on the way “ordinary” people employ moral terms (although this model of “ordinariness” often appears to be strongly informed by the philosophers’ own experience, enculturation and intuitions about morality and moral discourse). Yet, when analysing ordinary moral language to get at the meaning of moral terms, no clear or unproblematic picture emerges. As discussed above, there are multiple possible interpretations of the meaning of moral terms and how they might be cashed out metaphysically, and many of these interpretations are at odds. The end result is a Gordian knot of competing and largely mutually exclusive views, all purporting to best explain how we employ ethical terms, yet remaining tangled in counter-arguments, or demanding such a revision of our ethical language that they no longer reflect “ordinary” usage. Metaethics appears to be in something of a bind with no clear indication of how it might be untied, as Michael Smith sums up rather neatly:

> Nor should it be thought that this account of the deep disagreement that exists is misleading; that though there are disagreements, there are certain dominant views. The situation is quite otherwise. There are no dominant views... The scene is so diverse that we must wonder at the assumption that these theorists are talking about the same thing. (M. Smith, 1994)

I would suggest that it is the inside-out view's approach of starting with moral language in order to get at the essence of morality itself that is the central font from which this contention and confusion springs. By taking this approach, the inside-out view regards morality not as a natural phenomenon that emerges from the behaviour of creatures like us, but as an abstract conceptual framework, one that can be best understood by reflection inward rather than observation outward. As such, the success of the inside-out view is dependent upon there being a coherent concept that underlies our moral discourse in the first place. If it turns out that our ordinary moral discourse is in any substantial sense confused or inconsistent, that the concepts underlying it are fuzzy or vague, or that our moral language is influenced by our cultural predilections rather than referencing some metaphysical reality, then we might analyse everyday moral language all day long and yet fail to arrive at any concrete conclusion as to the unequivocal meaning of moral terms.

Some, such as Don Loeb, believe that our everyday moral discourse *is* indeed confused and inconsistent, to the point of being incoherent, thus suggesting that if we can draw any metaphysical conclusion, it would be a kind of anti-realism about morality at large:

> If both objectivity and its denial really are central and persistent features of our moral thought and talk, that would have profound implications for the debate over moral realism. Specifically, a form of irrealism would emerge. For if no adequate, coherent moral semantics can be found, then once again there doesn't seem to be anything (anything logically possible, anyway) to be realist *about*. More precisely, nothing we can be realists about would be entitled to unqualifiedly go by the name "morality," and so with at least many other terms in the moral vocabulary, at least insofar as they build in this incoherence. Perhaps we *can* pull a metaphysical rabbit out of a semantic hat in something like the way Ayer tried to. (Loeb, 2008)

Loeb argues that at least some of our everyday moral discourse employs both cognitivist and non-cognitivist elements in an unsystematic way, suggesting that ordinary moral discourse is incoherent. Cognitivism captures the notion that there are objective moral facts that provide a justificatory foundation to our moral beliefs. Non-cognitivism captures the notion that moral claims are evaluative and are intended to direct and motivate behaviour. Yet these two views are generally considered mutually exclusive because propositions are truth-apt but expressions are not, so moral utterances cannot be both.

Walter Sinnott-Armstrong (2009) gives a slightly different prognosis for the inside-out perspective and metaethics given the apparent inconsistency of our ordinary moral discourse. He suggests that our moral language might be "indeterminate", in that it can be equally well analysed in either cognitivist/realist or expressivist/anti-realist ways. If this turns out to be the case, Sinnott-Armstrong advocates a form of "Pyrrhonism": we should suspend belief about the truth of realism or anti-realism until such time as there is sufficient evidence or reason to choose between them.

Both Loeb and Sinnott-Armstrong are describing a serious problem for the inside-out approach to understanding morality. The problem comes down to the apparent fact that our everyday moral language is far from transparent, and is used in many ways by different people or in different contexts. It may even be inherently inconsistent or confused. To the extent that this is the case, then the ambitions of the inside-out programme are frustrated from the outset.

This does not necessarily mean that the inside-out programme is doomed, or that it might not yet yield insights into moral semantics or metaethics. Yet it does bode ill for the approach and its capacity to shed light on the nature of morality or to provide an explanation for phenomena such as moral diversity, particularly when that diversity is manifest in discourse. The inside-out metaethical perspective's narrow focus on moral language and the meaning of ethical terms has also led to a kind of philosophical myopia, whereby a great number of moral philosophers, particularly through the 20th century, ceased investigating morality as a natural phenomenon and suspended deliberation on how and why people behave in the ways they do in order to focus on resolving technical disputes about the possible interpretations of ordinary moral discourse and the meaning of ethical terms (although there are notable exceptions, such as John Dewey, who is quoted at the opening of this chapter).

4.4: Morality outside-in

The second approach to understanding morality takes the opposite view to the inside-out perspective. Instead of starting with the way we think and talk about morality working outwards to help explain moral phenomena, it starts by analysing patterns of altruistic and norm-driven behaviour among humans and works its way back *in* to help explain the way we think and talk about these behaviours. To the extent that such altruistic and norm-driven behaviours are relevant to morality, and are the subject of our moral utterances, then understanding the causes and dynamics of these behaviours could help shed light on the nature of morality – maybe even on our usage of moral terms. Given the direction of explanation, I call this approach the "outside-in" perspective on morality.

As mentioned in chapter 1, look at the world from an inside-out perspective and you see a landscape populated with agents scrambling to articulate justifications for their judgements of approbation or disapprobation about the actions of others. Yet, look at the world from the outside-in perspective and you see a landscape populated by organisms, such as us, engaging in complex social interactions and regulating their own and others' behaviour according to psychological impulses or norms, often in order to promote social and cooperative interaction. Both perspectives are reflecting on overlapping phenomena, but are addressing them from different points of view. This does not imply the two perspectives are contradictory; they may well prove complimentary. However, I would suggest that the outside-in view on morality could help resolve some inside-out questions, such as by explaining how we come to think and talk about morality the way we do, why moral norms are often considered to be binding and overriding, and why we might sometimes be confused or misguided when we employ particular moral terms.

One starting point for the outside-in perspective on morality – the starting point I choose for moral ecology – is to look at humans as biological organisms that are the product of millions of years of evolution. Like any other organism, humans have biological and reproductive "interests" that influence behaviour both directly and indirectly. And like any other evolved organisms, one would expect them to pursue their reproductive interests. After all, each organism alive today is the progeny of a long line of organisms that survived to reproduce in the past and, according to the strictures of the evolutionary process, these successful organisms tend to pass on their successful tendencies to their offspring.

Yet, as mentioned above, humans (and, to a lesser extent, some other animals) exhibit many curious behaviours, particularly in a social context. Just some of these include helping others even in cases when the helping activity imposes a cost on the helper (including a reproductive cost), the punishment of individuals who cause harm to others, the creation of behavioural rules that encourage such helping and punishing, and many other behaviours that appear to contribute to fostering a coordinated and cooperative social existence, even in light of the dangers that cooperation poses for one's reproductive interests.

4.4.1: Ultimate and proximate interests

It is worth taking a moment to clarify the term "interests" in the context of moral ecology. Biologically speaking, one can say that all organisms, including ourselves, are self-interested, to the extent they are typically motivated to pursue their biological interests. After all, organisms that prove ineffective at surviving and reproducing see their genes diminish from the gene pool. In this sense, an organism's interests refer to its biological imperative to reproduce. From this perspective, one might expect biological self-interest to manifest as behavioural self-interest. However, this is not necessarily the case. As mentioned above, we readily observe other-interested behaviour – including behaviour that appears to impose a reproductive cost on the actor – in a variety of organisms, not least in us. It would appear that, at least on some level, this behaviour is also motivated by the organism's interests, such as a genuine concern or empathy for another individual and a desire to aid them. How, then, to reconcile these two facets of "interest"?

The solution comes in distinguishing between at least two senses of "interest," and articulating how they interrelate. These two senses broadly correspond to the distinction between ultimate and proximate levels of explanation, as described by Ernst Mayr in "Cause and effect in biology" (Mayr, 1961). Ultimate explanations refer to *why* certain traits exist in an evolutionary context, where proximate explanations refer to *how* they operate on a more fine-grained level. One might ask, for example, why a particular bird chose to fly south at a particular time. One explanation could focus on the seasonal fluctuations in food sources and the benefits of migrating to more clement climes on the individual's chance of survival and, hence, its reproductive fitness. This is the ultimate level of explanation, and can offer an account of why the bird flew south, but not how. One could also cite the specific mechanisms that triggered the flight, such as the bird sensing a drop in temperature or changes in daily light levels. This is the proximate level of explanation, or the *how* rather than the *why*.

Likewise, one can view an individual's interests from these two levels of explanation. On one level are what we can call its *ultimate* interests, which represent the biological imperative instilled by evolution to survive and reproduce, and pass genes on to the next generation[6]. An organism that fails to serve its biological interests will likely see its genes eliminated from the gene pool in short order. As such, the organisms alive today – and their genes – tend to be the ones that have been highly effective at promoting behaviour that serves their ultimate interests.

However, ultimate interests are often opaque – or at best translucent – to the organism itself: the bird is unaware of the fitness benefits of various environments and climates. Rather, organisms tend to rely on a number of sensory mechanisms that have been shaped

[6] Biological interests can be viewed at the level of the individual or the gene. This can have important ramifications for explaining altruism towards immediate kin, for example (Dawkins, 2006; Hamilton, 1963), although this complication will be kept on the sidelines for the most part in this thesis as it does not significantly impact the broad thesis of moral ecology.

by evolution to gauge the state of their own biological and reproductive wellbeing as well as the state of the environment. They also have a suite of cognitive mechanisms and heuristics to interpret that sensory information and motivate behaviour in a way that tends to serve their ultimate interests (Godfrey-Smith, 1998). One example is the experience of pain and pleasure. Pain is typically triggered by damage or injury, causing an aversive reaction towards the stimulus that caused that damage or injury. Pleasure works in an opposite manner, causing an attract response towards stimuli that tend to provide some benefit to the organism.

Thus we also have what can be called an individual's *proximate* interests. These represent the particular psychological or physiological states that actually motivate behaviour, including beliefs and desires, both conscious and unconscious. The desire to avoid pain or seek pleasure can be considered proximate interests that guide behaviour moment-to-moment. Yet the mechanisms that produce these proximate interests have also been shaped by evolution to ultimately serve the individual's ultimate interests. Seeking a mate, for example, is crucial to facilitating reproduction for many species, including our own. Yet it is implausible that thoughts of gene frequencies in the population are front-of-mind when the lights dim, the bottle of red is uncorked and the Barry White LP is spun up. Individuals in this situation are almost certainly not contemplating the genetic compatibility of their partner's immune system when offering a compliment on their hair. And should the individual recount their experience to a close friend the next day, they're unlikely to cite a desire to increase the frequency of their alleles in the broader population as the reason they engaged in the behaviour they did. In fact, any human who actively steers their behaviour explicitly by considerations of their ultimate interests, such as citing the fitness advantages of procreation, would likely be considered a deviant. And if males were, in fact, motivated by reflection on genetic considerations, one might expect more men to donate sperm, or even pay for the privilege rather than requiring some payment as an incentive to spread their genes more widely for so little personal cost.

Of course, no sensory system is foolproof, and heuristics are rules of thumb that are prone to misfiring. This means proximate interests can often fail to serve ultimate interests. Our penchant for sweet and fatty foods, for example, evolved in a time when energy dense foods were rare and highly prized. A strong reward mechanism that encouraged us to seek out such energy dense foods proved to be adaptive. However, this is not to suggest that such proximate interests are still adaptive today, particularly in an environment filled with energy dense sweet and fatty treats; overconsumption of these can ultimately harm our health and our biological interests (Birch, 1999). Yet, over the tides of evolutionary time, natural selection has generally shaped the mechanisms that underpin proximate interests to serve the ultimate interests of the organism and its genes.

This distinction between proximate and ultimate interests is particularly important in the moral context because of a tendency to conflate the various levels of interests. This can be seen in Michael Ghiselin's memorable statement: "scratch an 'altruist,' and watch a 'hypocrite' bleed" (Ghiselin, 1974). Ghiselin is referring to cases when an individual engages in some apparently altruistic costly helping behaviour, yet in doing so they benefit from the fruits of cooperation or reciprocation, thus furthering their ultimate interests. What on the surface appears to be altruistic is, in fact, self-interested behaviour. However, this conflates the two levels of interest. Just because an example of costly helping might tend to enhance the fitness of individuals who perform it does not mean that they are not genuinely psychologically motivated by concern for the interests of others rather than concern for their own perceived interests. As such, it is not necessarily contradictory to suggest that genuine (proximate) psychological altruism exists even if it turns out to advance (ultimate) self-interest.

Philip Kitcher carefully defines "altruistic behaviour" in light of this possible confusion as being behaviour that *appears* to be helpful, but which *may or may not* be motivated by (proximate) self-interest, or "Machiavellian" intentions, such as an expectation of reciprocation. Thus the behaviour is "helpful" to another individual regardless of the motivation behind it. He contrasts this with "psychological altruism", which is motivated by genuine magnanimity and feelings of concern for others:

> When you come to see that what you do will affect other people, the wants you have, the emotions you feel, the intentions you form, change from what they would have been in the absence of that recognition. (Kitcher, 2011)

Thus one can now more clearly see the origin of Ghiselin's confusion. If an individual engages in some "altruistic behaviour" with an expectation of receiving a greater benefit in the future, one might well call them a hypocrite. However, if they are genuinely motivated by feelings of concern, and they hold no explicit expectation of reciprocation, then they are in a sense acting genuinely psychologically altruistically. It might well turn out that such proximate altruism *does* tend to advance an individual's ultimate interests – and that we have evolved psychological mechanisms that promote such psychological altruism (Haidt & Kesebir, 2010; Hauser, 2006; Joyce, 2006) – but that does not make such altruistic behaviour hypocritical in Ghiselin's sense.

Thus, there are at least two distinct senses of "interest," the first being ultimate interests, such as survival and reproduction, and the second being proximate interests, including the psychological beliefs and desires that motivate behaviour. In the remainder of this thesis I will attempt to clearly render interest in its appropriate sense given the context.

4.5: Moral phenomena

It is the helping behaviour that many organisms – particularly *Homo sapiens* – engage in, and the existence of the proximate mechanisms that promote such behaviour, that is of interest in this thesis, particularly in terms of how those mechanisms can produce diversity in moral attitudes, norms and behaviour. This helping behaviour also appears to be closely linked to much of the usage of ordinary moral language mentioned in section 4.4 above, such as in the way we judge others' behaviour, compel them to act in certain ways or justify the norms we create and obey.

The outside-in approach to morality takes such phenomena as its starting point of enquiry. Yet it does not just limit itself to looking at individual behaviour, but is also interested in the patterns and dynamics of behaviour as they are observed at the population level over a single or many generations. This is in the same vein as a behavioural ecologist might take the behavioural patterns of some species within a particular environment as the starting point of her explanation for those behaviours. Or as a sociologist might take observations of institutions in action as the starting point of his explanation for their existence and origins, rather than solely the beliefs of those who interact with those institutions.

The outside-in view on morality that I take is largely an empirical endeavour. It begins by observing moral phenomena, including social and cooperative interaction, altruistic and helping behaviour, the creation and perpetuation of social and moral norms, and the patterns of moral discourse, and then goes on to postulate explanations for those phenomena. As such, it draws upon a range of empirical disciplines to help understand and explain moral phenomena, including anthropology, psychology, behavioural ecology, game theory and evolutionary biology. It is also a thoroughly naturalistic approach to morality. It is entirely sympathetic with a metaphysically or scientifically naturalistic worldview: one that believes only natural forces or things exist, and one opposed both to supernaturalism and non-naturalism.

This outside-in view of morality is not entirely new. Others have pursued a similar approach to investigating moral phenomena under a variety of banners. Around a century ago pragmatists such as John Dewey argued that morality is a natural phenomenon and ought to be studied as such. In the late 20th century, sociobiologists such as Edward O. Wilson (1980) and Richard Alexander (1987) advocated a kind of grand synthesis of biology and ethics, supported by philosophers of biology such as Michael Ruse (Ruse & Wilson, 1986). More recently, some moral psychologists and philosophers such as Jonathan Haidt and Jesse Graham (Haidt & Graham, 2009; Haidt & Kesebir, 2010) have advocated a broadening of our perspective on morality to include a wider range of social and behavioural phenomena, such as community, divinity and purity rather than just harm and fairness – a position that is in sympathy with the outside-in approach. The philosophers Owen Flanagan, Hagop Sarkissian and David Wong (2008) have promoted an approach to ethics they call "human ecology", a science "concerned with saying what contributes to the well-being of humans, human groups, and human individuals in particular natural and social environments" that is along very similar lines to the outside-in perspective.

The philosopher Joshua Greene (2002) has also made a distinction between two senses of the term "moral" that is reminiscent of my inside-out/outside-in distinction. First is what he calls "moral$_1$", which is "of or relating to the facts concerning right and wrong, etc." Here he is referencing not only a view of morality as being grounded in how we think and deliberate about matters of right and wrong, but specifically the objectivist moral realism that commonly emerges from this perspective. Contrasting this is "moral$_2$", which he defines as "of or relating to serving (or refraining from undermining) the interests of others." By this he is referring to the same kind of moral phenomena mentioned above rather than solely moral discourse, much as does the outside-in perspective. Greene claims that both senses of the term "moral" occur in ordinary moral discourse, and argues that it is a mistake to accept a realist explanation of such discourse when the alternate sense suggests a different interpretation of what morality is, and one that is not dependent on metaphysically problematic "moral facts." While Greene's distinction is not identical to the inside-out/outside-in distinction, particularly as Greene is still referring to how we should understanding the *meaning* of the term "moral," his usage of the distinction can be read as favouring an outside-in approach to understanding moral phenomena.

However, despite these pockets of alternate outside-in-like views on the subject and methodology of ethics, it appears the majority of philosophers writing on the nature of morality still take the inside-out approach as exemplified by much contemporary metaethics. As such, one of the first hurdles to clear in advocating a shift in perspective on morality is to overcome the inertia inherent in viewing morality as being a special domain of action that is largely divorced from more earthly concerns, such as cooperation. This delineation of the moral domain from the outside-in perspective will be the subject of the next chapter.

Chapter 5: The Moral Domain

> I must help other people, and do everything I could for other people, and look out for them all the time, and never think about myself… I went out in the woods and turned it over in my mind a long time, but I couldn't see no advantage about it – except for the other people.
> - Huckleberry Finn

5.0: From cooperation to morality

As discussed in the last chapter, the outside-in perspective on morality starts with the observation that many *Homo sapiens* tend to behave in interestingly prosocial and cooperative ways, and focuses on the psychological and cultural mechanisms that promote such behaviour. This thesis has a further interest in how these mechanisms produce variation and diversity in norms and attitudes that promote this prosocial and cooperative behaviour. This emphasis on prosocial and cooperative behaviour might inspire an observation that this thesis is ultimately concerned with *cooperation* rather than *morality*. Indeed, much of this thesis discusses the dynamics and complexities involved in promoting cooperation amongst self-interested agents. However, I maintain that it still has something to say about morality.

Naturally, the definition of morality is hotly contested, and I have no intention to weigh in and offer a definition that will satisfy everyone. Rather, in this chapter I hope to offer some observations on how morality is often conceived from the inside-out perspective, and contrast that with an alternate view from the outside-in perspective. Specifically, I am interested in exploring the bounds of the moral domain, presuming such a domain exists. I will argue that the inside-out perspective tends to narrow the moral domain down to a collection of norms that possess a special binding authority and concern matters of harm and fairness. I will then offer an alternative delineation of the moral domain from the outside-in perspective that places more emphasis on interestingly prosocial and cooperative behaviour, and the norms and attitudes that promote them. Like the inside-out/outside-in distinction, this is best seen as a rendering of alternate perspectives on the subject of morality rather than mutually exclusive positions. Yet to the degree that the outside-in perspective can shed light on the phenomenon of moral diversity as described in chapter 2, then it may be of some use to philosophers. The sketch of the moral domain from the outside-in perspective will then lead to a functionalist definition of morality, which will be the subject of the next chapter.

5.1: The moral domain inside-out

Typical of the inside-out approach to morality is to analyse ordinary language in an attempt to establish what qualifies a moral rather than a non-moral utterance. As discussed in chapter 3, there are many examples of inside-out analyses of moral language, from G. E. Moore's reflection on the meaning of "the good" (Moore, 1903), to A. J. Ayer's expressivist interpretation of moral assertions (Ayer, 1936), to John Mackie's concern that the facts apparently referenced by moral utterances do not exist (Mackie, 1977) – and countless other examples besides. Rather than compare them all, I will draw primarily on the analysis given by Richard Joyce in *The Evolution of Morality* (2006) as being representative of the inside-out view on morality. To the extent that Joyce's analysis resembles other inside-out views, then it will form a foundation against which to contrast the outside-in view on the moral domain. I will also draw on the work of Eliot Turiel as being representative of a strong empirical tradition that has explored the bounds of the moral domain.

Joyce is particularly concerned with questions such as whether morality is somehow "innate," whether morality "is ultimately something that helped our ancestors make babies" as well as the question of whether evolution undermines certain metaethical theories, such as realism. After establishing that there is compelling empirical evidence supporting the notion that humans have evolved to be "helpful" – if in a limited and conditional sense – he goes on to stress that there is a profound difference between helpfulness and *morality*:

> I want to establish… that in kin selection we have a quick and easy, empirically supported, evolutionary explanation for why humans might have "prosocial emotions" (e.g., love) toward certain others: emotions that provide the motivation for helping behaviours. But what I really want to emphasize here is how far this answer falls short of explaining *morality*. (p.49)

Joyce then goes on to reflect on our typical usage of moral language to argue that morality must be considered as something more than just "helpfulness." He does acknowledge that an individual who is motivated purely by prosocial emotions is still worthy of praise, but only in a limited sense. For, according to Joyce, individuals who only have an emotional *inhibition* against causing harm, for example, are not moral in the same way as are individuals who subscribe to a *prohibition* against causing harm; and it is only the latter that concerns morality:

> My point is… that someone who acts solely from the motive of love or altruism *does not thereby make a moral judgment*… If, then, our object is to investigate whether it is part of human nature to make moral judgments – to think about each other and the world in moral terms – we must conclude that an explanation of how natural selection might end up making human beings with altruistic, sympathetic, loving tendencies toward each other – how, that is, it might produce *nice* humans (or, if you prefer, *virtuous* humans) – misses the target. (p.50-51)

Note Joyce's definition of what constitutes a moral judgement as being "to think about each other and the world in moral terms." He thus narrows down morality to exclude "nice" or "altruistic" behaviour motivated by emotion, instead honing in on the way we *think* (and perhaps *talk*) rather than the way we *act*. Whilst perhaps a radical position, I take Joyce's emphasis on thinking, talking and judgement to be characteristic of the inside-out approach to morality and to metaethics at large. Given Joyce's identification of morality as moral judgement, he then picks out two key features of most moral judgements. The first concerns their normative form, the second the typical subject matter of moral judgements.

5.1.1: Normative form

Joyce points out that moral judgements tend to possess a kind of "inescapable authority" that is not found in many other kinds of judgements. This is in the sense that a moral judgement does not seem to conditionally apply to an individual by virtue of their subjective preferences or perceived ends: it seems somehow wrong to excuse someone from a prohibition against torturing because they feel like torturing, or because they believe torturing is in their interests. In Kantian parlance, moral assertions are said to be *categorical* rather than *hypothetical* imperatives; the former binds an agent to acting irrespective of their desires, while the other is contingent on the agent's desired ends (Kant, 1785).

Joyce cashes out this apparently essential feature of morality in terms of both the "inescapability" and "authority" of moral judgements – the conjunction of which he terms "practical clout" – such that moral assertions appear to bind an individual to behaving a certain way (and justify their punishment if they refuse) irrespective of that individual's preferences or ends, or whether the individual even opts out of morality altogether. (It is worth noting that Joyce's observation that moral judgements appear to possess practical clout does not imply they actually do have such clout, only that most people tend to talk as if they do.)

This notion of the inescapability of moral judgements also featured prominently in the work of Eliot Turiel (1983), who built upon the prior research of Lawrence Kohlberg (Kohlberg & Hersh, 1977) and Jean Piaget (1932) exploring the development of moral judgement in children. Turiel conducted a series of landmark experiments in the late 1970s and early 1980s that purported to reveal an innate tendency to distinguish *moral* from *conventional* norms. Moral norms, he claimed, had an objective prescriptive force that was not contingent on any authority. They were perceived to be universally applicable rather than locally binding, and typically involved cases of harm, injustice or the violation of rights. Conventional norms, on the other hand, were perceived to be arbitrary, lacked objective prescriptive force, tended to apply locally rather than universally, and concerned matters of coordination and custom. According to Turiel, a breach of a moral norm is typically considered to be far more serious than the breach of a conventional norm, and they are sanctioned appropriately. Turiel found that even children were readily and intuitively able to distinguish between these two categories of norms. This ease of discrimination led some to postulate that the moral/conventional distinction might be innate rather than being a learned convention itself (Dwyer, 2009) – although it is worth stressing that Turiel himself stated that the distinction is likely constructed during enculturation and development.

5.1.2: Subject matter

Where many inside-out analyses of moral language focus exclusively on the normative form of moral utterances and judgments, Joyce also turns his attention to the subject matter appropriate to such judgments. After all, categorical judgements possessing practical clout could, in principle, apply to just about anything – even the act of looking at hedgehogs by the light of the moon. If we decide that such a thing is simply not a valid moral concern, then we need to establish some kind of boundary around moral domain. The question is whether there is any particular subject matter, or domain of concerns, that are suitable subjects for moral judgements.

Joyce first rejects the notion that moral judgements must include any particular proscriptions, such as against infanticide, as this would obviate any possibility of moral debate over those proscriptions. He then considers what broad types of proscriptions tend to qualify as moral. He reflects on the kinds of things to which moral predicates are typically applied in ordinary moral discourse and finds a number of apparent universals:

> (1) negative appraisals of certain acts of harming others, (2) values pertaining to reciprocity and fairness, (3) requirements concerning behaving in a manner befitting one's status relative to a social hierarchy, (4) regulations clustering around bodily matters (such as menstruation, food, bathing, sex, and the handling of corpses) generally dominated by concepts of purity and pollution. (p.65)

He then makes the observation that much of this list – the first three points in particular – can be digested down into "prescriptions and values that seem designed to protect and sustain social order, to resolve interpersonal conflicts, and to combat the rampant pursuit of individual welfare... in particular, a great deal of the moral domain is devoted to matters pertaining to how humans may harm each other." The fourth category he considers somewhat problematic because it appears to concern self-regarding rather than other-regarding actions, but he still asserts that the bulk of any moral system is likely to concern interpersonal relations, with a particular emphasis on preventing harm and promoting fairness. It is also worth noting that Joyce qualifies these as being subjects of moral systems not because they represent interestingly social or cooperative phenomena in their own right, but only to the extent that they are the kinds of things to which *moral judgements* with practical clout typically apply.

Turiel also saw the moral domain as being limited to concerns of justice, rights and welfare, with issues outside of these areas diminished to mere convention. His own experiments, which were exclusively conducted on children, focused on highlighting this distinction by offering paragon – typically "schoolyard" – examples of harm and injustice as moral norms and examples of custom as conventional norms. Even when Turiel's experiments were adapted for other contexts, such as exploring the moral judgements of psychopaths, they retained their simplistic "schoolyard" character rather than exploring any grey areas that might exist between the moral and conventional (Kelly, Stich, Haley, Eng, & Fessler, 2007). If Turiel's definition of the moral domain as concerning exclusively matters of justice, rights and welfare holds, then from this perspective moral diversity reflects diversity only in individuals' views on justice, rights and welfare, and diversity in conventional norms is primarily due to their arbitrary, contingent, authority-led or subjective nature.

5.1.3: Narrow morality

The inside-out view of the moral domain, as exemplified by the work of Joyce and Turiel, tends to see morality as being defined in terms of a special kind of judgement about norms concerning harm and fairness. This reflects the emphasis placed on moral language and discourse by the inside-out perspective, particularly influenced by the kinds of discourse individuals in contemporary Western societies engage in. This means that in order for an action or concern to be considered moral, it needs to qualify on two broad accounts. The first is that it includes some form of moral judgement, meaning some reference to a categorical norm that carries practical clout. If an individual behaves in an other-interested (i.e. "helpful"), altruistic or otherwise prosocial or cooperative way, but that behaviour is not motivated by obligation to a binding norm, then it is not strictly speaking moral. The second feature that qualifies an act or judgement as being moral is that it primarily concerns preventing harm or promoting fairness. If a judgement is about a binding norm concerning some other subject, such as a manner of dress, a food taboo or some act that does not impact others at all, then it is often not considered a moral judgement.

Interestingly, this distinction between normative form and the subject matter of morality might help to explain an unusual phenomenon identified by Jonathan Haidt, Fredrik Björklund and Scott Murphy (2000), which they call "moral dumbfounding." In their study, they presented 30 undergraduates at the University of Virginia a number of vignettes of actions ranging from a classic moral dilemma from Lawrence Kohlberg's playbook to cases of incest and cannibalism. Crucially, the latter cases were constructed such that no-one involved experienced any harm. The subjects were asked whether the acts depicted in the vignettes were wrong, and if so, why. Those who identified the harmless acts of incest and cannibalism as wrong (28 per cent and 32 per cent, respectively) often had difficulty giving any firm reasons *why* it was wrong. They often gave what Haidt et al. call "unsupported declarations", such as: "It's just wrong to do that". They also appeared "dumbfounded", where they "thought an action was wrong but they could not find the words to explain themselves". Drawing on the discussion above about the two features of common moral discourse it is possible to conjecture an explanation for this phenomenon of moral dumbfounding. It could be that the "dumbfounding" situations involved some proscription that possesses the normative form of typical moral judgements – i.e. they were seen to possess inescapable authority – yet they fell outside the presumed subject matter of morality because they did not inflict harm or cause an injustice. The normative form helped trigger a moral reaction, yet reflection failed to find a justification, hence the "dumbfounding."

Due to the moral domain being restricted to binding norms that concern harm and fairness, the inside-out perspective sees morality in fairly narrow terms. Behaviour that is motivated purely by a magnanimous spirit, virtuous training, prudential concerns or by self-interest, no matter how helpful it is to others, is not *really* moral. Likewise, behaviour that is motivated by what Kant would call a hypothetical imperative, no matter how helpful to others, is not *really* moral. Furthermore, behaviour that some individuals and cultures would likely call moral, such as food taboos or norms concerning appropriate dress, are also not *really* moral. From this perspective, the idea that morality is primarily about cooperation is not tenable, unless that cooperation is promoted via categorical norms concerning harm or fairness.

5.2: The moral domain outside-in

As discussed in the last chapter, the outside-in perspective on morality is less concerned with what people tend to think or say about morality, and more with how they act in interestingly helpful and cooperative ways, and what prompts them to behave in such ways. As such, the outside-in perspective leads to a somewhat different take on the bounds of the moral domain.

For a start, if only behaviour that is motivated by a commitment to binding norms concerning harm and fairness is considered moral, then this excludes much of the helpful, cooperative and prosocial behaviour that we observe in our species from qualifying as being moral. Yet even in everyday discourse I would suggest that behaviours motivated purely out of a sense of magnanimity, or motivated by emotions such as empathy or loyalty, would often be considered moral. For example, cultivating empathy appears to be one of the pillars of moral education in many cultures around the world, along with encouraging altruistic behaviour motivated by that sense of empathy. Philosophers from David Hume (1739) to Jesse Prinz (2007) have considered sentiments such as empathy to be central to motivating moral behaviour. Recent research by Jonathan Haidt has also highlighted the importance of emotions such as empathy in motivating altruistic behaviour, and has suggested that rational justifications of such actions are often post-hoc rationalisations of the emotion-driven action (Haidt, 2001). Granting that Joyce believes that moral judgements do possess a conative component along with some belief or judgement, his conception of morality as requiring some sense of obligation or prohibition means that many interestingly prosocial and cooperative behaviours would fall outside of the moral domain.

There is certainly something very interesting about norm-driven behaviour that suppresses self-interest and encourages altruistic behaviour, particularly in our species, as

I will discuss in later chapters. However, it would appear that norm-driven behaviour is only one facet of morality. As will be discussed in more detail in chapters 6 and 8, the outside-in perspective on morality sees moral norms as a key component of culture that has helped to solve many of the problems inherent in social living (Kitcher, 2011). Moral norms are considered to be a subset of behavioural norms, which in turn are guides that promote particular behavioural strategies in specific contexts. Norms alter the behavioural traits of those who conform to them, not only through an explicit rational calculus, or an aversion towards punishment, but also through being internalised and altering the individual's behavioural dispositions. It is an interesting observation that many of the norms that appear to concern promoting prosocial and cooperative behaviour – and inhibiting self-interested and socially disruptive behaviour – tend to be rendered as binding and carrying practical clout. But there are also many other norms and conventions that appear to promote prosocial and cooperative behaviour that are not rendered as categorical imperatives. Many of the norms and customs that promote group conformity, for example, such as dress codes or basic etiquette, might not be considered to be moral in a moral/conventional task such as those set by Turiel. However, they do represent interesting cultural devices that appear to help facilitate social living and altruistic behaviour. To that extent, they fall within the moral domain from the outside-in perspective.

5.2.1: Subject of morality

It does seem to be the case that most educated Westerners would consider the core subject matter of morality to be the prevention of harm and the promotion of fairness. However, if harm and fairness are the central concerns of morality, what to make of the many other individuals in cultures that apply apparently binding norms with practical clout to concerns outside these domains? For example, there are many cultures that consider matters of purity and sanctity to be suitable topics inspiring moral judgement (Graham et al., 2011; Haidt, Rosenberg, & Hom, 2001; Richard A. Shweder, Much, Mahapatra, & Park, 1997).

Furthermore, Turiel's results discussed above do not necessarily imply that there indeed exists clean and universal distinction between subjects that fall into the moral rather than the conventional domains. Turiel's experimental results certainly appear to highlight an important distinction between how many people address these two classes of norms, particularly in response to breaches thereof, but there is still a question as to whether the moral domain is limited only to issues of harm, rights and welfare. It might just be the case that breaches of moral norms are considered to be significantly more serious than breaches of conventional norms, but this does not necessarily fix what the content of moral norms must be. In fact, recent research by Jonathan Haidt and colleagues has shown that some people – particularly those of lower socioeconomic class – tend to treat some concerns outside of harm, rights and justice as being characteristically moral transgressions, even if they are unable to identify any harm or rights violation that has occurred (R.A. Shweder & Haidt, 1993). Other research by Haidt and colleagues (Haidt et al., 2000) has shown that actions inducing a sense of disgust are often seen by American children (the subjects of Turiel's studies) as moral transgressions without them concerning issues of justice, rights or welfare. Such cases offer more instances of "moral dumbfounding," as discussed above. It also appears that some actions involving harm may *not* trigger a moral response (Kelly et al., 2007).

As such, while it does seem that there are at least two classes of *response* to transgression of norms that treat such transgressions with varying degrees of severity, it is not clear that there is any universally realised class of issues or actions that quality as *moral* as opposed to *non-moral*. It might instead be the case that the "moral" response is often reserved for those transgressions that are considered to be the most serious within a particular culture. And it might be the case that most cultures consider transgressions of justice, rights or welfare as the most important. Yet there does not appear to be any necessary link between moral response and justice, rights and welfare.

5.3: Expanding the moral domain

As mentioned above, it has been a popular view in ethics, metaethics and moral psychology over the past several decades to view the moral domain as being primarily concerned with issues of justice, rights and welfare, or as Haidt (2007) compresses it, "harm and fairness". However, this may be more of an artefact of history and of the particular culture and lifestyle of those living in modern post-industrial society than being indicative of any necessary class of moral concerns. In fact, evidence cited above suggests individuals from pre-industrial cultures – either in the West prior to the industrial revolution or in modern cultures that have yet to be significantly reformed by modernisation – often consider the moral domain to be considerably broader than just concerning harm and fairness.

So why the recent penchant for focusing on harm and fairness? Jonathan Haidt and Selin Kesebir argue that morality went through what they call the "great narrowing" during the

Enlightenment, shrinking the bounds of the moral domain from the broad community- and virtue-based ethics of the ancient and traditional world (and much of the non-developed world to this day) to the rationalistic and deliberative ethics of the Enlightenment (Haidt & Kesebir, 2010). While harm and fairness do appear to be nearly universally represented in moral systems around the world and throughout history, many cultures have not limited their moral concerns to just those two features. Many didn't even consider there to be a separate domain of moral concerns that was distinct from other features of their social existence. In ancient and "traditional" societies morality was typically interwoven with many other aspects of culture, including broad social norms, religious practices and customs. However, during the Enlightenment, philosophers surgically separated morality into a separate domain, with a degree of independence from other aspects of culture.

One reason for this may have been the growing importance of social institutions, such as the judiciary, which took over some of the functions previously handled on a more individual-to-individual level by social and moral norms. Edward Westermarck gives examples of cultures that adopt a form of independent arbitration of disputes, and the effect this has on reducing rates of retributive violence (Westermarck, 1906). Presumably, over time, such an institution of arbitration – if managed in a trusted and uncorrupted manner – could see the social norms that demand direct retribution for perceived wrongs atrophy. More recently, Jared Diamond gives an example of how the introduction of even a small number of modern police, armed with firearms, into Papua New Guinea was able to significantly reduce the frequency and scale of tribal warfare (Diamond, 2012). The New Guineans embraced a new institution that was able to resolve their disputes with less violence, and thus often deferred to this institution rather than pursuing justice via their traditional customs and norms of revenge.

Where these institutions emerge, particularly post-Enlightenment institutions, they are often backed by explicit rationalisation of their purpose and function, and such rationalisation requires reflection on the ends of those institutions, such as justice and the prevention of harm. It is this process of rationalisation and justification that may account for the greater emphasis on harm and fairness in cultures that have a greater number of institutions taking over the role of other social and moral norms, as these issues are the ones being actively debated, and thus recognised as being moral.

Another reason for the "great narrowing" could be the relationship that has existed between religion and morality, and particularly the Judeo-Christian influence on Western philosophy and ethical thinking during the Enlightenment. Haidt and Kesebir point out that the 18th century was a tumultuous and eventful time for religion and philosophy both. Religion in Europe was embattled in the political sphere, and the perceived influence of the divine in turning the cogs of the natural world was diminishing. During this time European philosophers – particularly the scientifically-minded "natural" philosophers – were attempting to render a thoroughly naturalistic picture of the world absent of any supernatural forces. Likewise, many ethicists were attempting to reconstruct morality and ground it in forces other than the divine. A paragon example is Immanuel Kant, who sought to ground morality in the strictures of reason alone. Yet even he produced a morality that bears a striking resemblance to the internalised binding rules that underlie the Abrahamic religions, only with reason replacing God as the author and motivator of our moral impulses[7].

Perhaps unsurprisingly in an age when reason and reflection were becoming the preferred tools for reckoning the foundations to the natural world, they were also becoming the primary tools for reckoning the foundations of morality, which focused on how we might justify our judgements about how ourselves and others ought to behave. The two dominant traditions that emerged from Enlightenment ethical thinking, deontology and consequentialism, turned morality from an issue of behaving in accord with communal norms or with divine command into being an issue of behaving in accord with abstract rational principles primarily concerning preventing harm and promoting fairness. As Haidt and Kesebir state:

> Deontologists and consequentialists have both shrunk the scope of ethical inquiry from the virtue ethicist's question of "whom should I become?" down to the narrower question of "what is the right thing to do?" The philosopher Edmund Pincoffs (1986) documents and laments this turn to what he calls "quandary ethics." He says that modern textbooks present ethics as a set of tools for resolving dilemmas, which encourages explicit rule-based thinking. (op. cit.)

They do stress that this defining of the moral domain is descriptive rather than prescriptive; they are not suggesting that we *ought* to set the bounds of the moral domain at any particular point, only that there are many individuals and cultures around the world and throughout history who have considered the ethical domain to be broader than that rendered by many Enlightenment and post-industrial thinkers.

[7] Even then, Kant's moral philosophy is not entirely secular, and involves some references to God.

Owen Flanagan, Hagop Sarkissian and David Wong also conjecture that the popularity of realist and non-naturalist accounts of morality, particularly from North American philosophers, in the last century or so is another throwback to the Christian underpinnings of North American culture (Flanagan et al., 2008). The pervasiveness of a divine command moral worldview has encouraged a reflection on morality as concerning binding imperatives, with a tendency for philosophers to attempt to replace divinity with reason as the justification for these imperatives. And, as much contemporary Christian morality concerns issues of harm and fairness, these moral foundations garner increased attention in today's moral musings.

Haidt and his colleagues have conducted further research showing there appear to be at least five foundations of morality that regularly appear in moral systems throughout history and across the world, although some cultures weigh the foundations differently. The five hypothesised foundations are

> 1. Harm/care: concerns for the suffering of others, including virtues of caring and compassion.
> 2. Fairness/reciprocity: concerns about unfair treatment, inequality, and more abstract notions of justice.
> 3. Ingroup/loyalty: concerns related to obligations of group membership, such as loyalty, self-sacrifice and vigilance against betrayal.
> 4. Authority/respect: concerns related to social order and the obligations of hierarchical relationships, such as obedience, respect, and proper role fulfilment.
> 5. Purity/sanctity: concerns about physical and spiritual contagion, including virtues of chastity, wholesomeness, and control of desires. (Haidt & Kesebir, 2010)

Why these particular foundations? Haidt and Kesebir have argued that these five foundations correspond to various *problems of social living* that our ancestors have had to solve over many millennia of biological and cultural evolution. While preventing harm and unfair distribution in the face of rash or self-interested motives continue to be serious problems to be solved to this day, so too were problems of enabling coordinated action, encouraging group cohesion and maintaining conformity to local norms in many less modern societies. If this is the case, then the bounds of morality have been – and still are in many corners of the globe – conceived in ways broader than just concerning harm and fairness.

5.4: Function of morality

How then to delineate the bounds of the moral domain? The inside-out view tends to regard morality in terms of moral judgement, with a corresponding interest in our usage of moral terms – particularly those apparently related to bindingly prescriptive moral norms that carry practical clout – and tends to see morality as principally concerning issues of harm and fairness. However, this approach risks dismissing many interesting phenomena that bear explanation, not least many features of moral diversity.

In contrast, the outside-in view sees morality as a natural phenomenon, one that is readily visible in the interestingly social and cooperative behaviour of creatures like us. It certainly does appear that we tend to employ moral terms loaded with a special binding prescriptive force, but this is just one of the interesting phenomena to be explained. It also seems as though issues of harm and fairness are widely considered to be important to morality. This too is an interesting phenomenon to be explained. However, humans also act in interestingly helpful and altruistic ways, prompted by prosocial emotions, habit, customs or prudential or conventional norms. We also appear to employ norms with binding prescriptive force to issues other than harm and care.

As such, the outside-in view on morality is not restricted to looking at only moral judgement or binding norms concerning issues of harm and fairness. Instead it looks at morality as being a broad and vaguely-bounded domain of issues that concerns the behaviours and norms that promote helpful, altruistic and cooperative behaviour. In this way, the outside-in perspective is less concerned by what people tend to think morality *is* rather than what it *does*. One notion that binds these various features of our helpful, altruistic and cooperative behaviour and norms together is the function that they serve in facilitating social living. So it is by looking at what function morality has had to play in our societies and cultures that we can better understand what kinds of concerns have typically been considered moral, and why the content of morality varies from one culture and from one point in history to the next. As such, it is to moral functionalism that we will turn in the next chapter.

Chapter 6: Moral Functionalism

> What they say is that it is according to nature a good thing to inflict wrong or injury, and a bad thing to suffer it, but that the disadvantages of suffering it exceed the advantages of inflicting it; after a taste of both, therefore, men decide that, as they can't evade the one and achieve the other, it will pay to make a compact with each other by which they forgo both. They accordingly proceed to make laws and mutual agreements, and what the law lays down they call lawful and right. That is the origin and nature of justice.
> - Glaucon in *The Republic*

6.0: Moral functionalism

In traditional Fijian culture it is taboo for a pregnant or lactating woman to eat certain seafoods. Judaism famously declares that creatures with cloven hoofs that do not chew their cud – notably pigs, camels, hares and, unexpectedly, the hyrax – are off the menu. In most Western countries there is an unwritten prohibition against eating dog meat. Food taboos such as these are commonly found in spectacular diversity in moral systems from across the world and throughout recorded history. The only thing perhaps more remarkable than the near ubiquity of food taboos is the diversity of foods that are considered taboo.

Food taboos present something of a quandary for the inside-out perspective on morality. On the one hand, prohibitions against eating certain types of food do not appear to prevent harm or injustice, yet they often carry the binding normative form of moral proscriptions. The justifications for these norms also vary widely, if they are even explicitly articulated at all. Some are maintained by tradition, as are the Fijian food taboos, others are given a supernatural foundation, such as those of the Judeo-Christian and Islamic ilk, and others have little or no explicit justification, such as the Western aversion towards eating dog meat (Haidt, Roller, & Dias, 1993; Haidt & Hersh, 2001). Yet the claim that food taboos simply are not "moral" is likely to fail to impress those who take such taboos seriously, and for whom the breach of such a taboo is considered as a serious breach of moral strictures – which would likely include a majority of humans who have lived over the past few thousand years, and a significant proportion of them alive today. As such, food taboos appear to qualify as a moral phenomenon that deserves an explanation, even if such an explanation is not to be found in the justifications offered by those who maintain such taboos.

So how is the existence of food taboos to be explained? Why do they tend to emerge and persist, even if they appear to be irrelevant to advancing causes of preventing harm or promoting fairness? How is the diversity of foods that are considered taboo to be explained? Where the inside-out perspective on morality might struggle to answer these questions – should it even resist the temptation to dismiss food taboos as non-moral – there is another perspective that can help to bring clarity to such moral phenomena: moral functionalism.

One consequence of the inside-out approach to morality is a focus on defining morality in what might be considered *essentialist* terms. Most philosophers have focused on questions of what morality *is* rather than what it *does*, namely the *content* rather than the *function* of morality. While it makes sense from an inside-out perspective on morality to speculate on whether our moral utterances reference some realm of special moral facts, from the outside-in perspective a more interesting question is to ask what impact morality has on the behaviour of individuals who employ it. In this respect it is reminiscent of Greene's moral$_1$/moral$_2$ distinction mentioned in section 4.5. However, before examining what the function of morality might be, it is worth clarifying in more detail what a *function* itself is.

6.1: Functions

Functional definitions might not be common in ethics, but they have a long pedigree in biology, and it is from this tradition that I draw in order to define morality in functionalist terms. We naturally talk about function for all manner of things: the function of a toaster is to turn bread into toast; the function of the heart is to pump blood; the function of the warbler's call is to attract a mate, etc. Such functional ascriptions tend to appeal to some kind of purpose or design or goals that the thing fulfils, i.e. we tend to think of function *teleologically*. This is rather unproblematic when it comes to artefacts such as toasters, which tend to have a designer who sets the intended purpose and shapes its features to satisfy that purpose. However, it is more problematic when talking about biological traits like hearts and bird calls because, unlike toasters, organisms do not happen have a designer to set the intended purpose of the organism or trait. In this context, it is inappropriate to employ explicitly teleological explanations analogous to those offered for artefacts when describing the function of biological traits. The solution comes from redefining the way we talk about function to make it *etiological* rather than teleological, i.e. backward-looking rather than forward-looking, in order to factor in natural selection.

Broadly speaking, where a teleological functionalist explanation would say X was *designed* to do Y, an etiological functionalist explanation would say X was *selected for* because it does Y. Thus an etiological account looks backwards and employs function to explain why the feature or trait exists, often with an appeal to natural selection to provide that backward-looking story. For example, we know a heart's function is to pump blood precisely because it is this activity that explains why the heart exists and why it exhibits the design and physical properties – i.e. the "morphology" – it has. Yet pumping blood is just one of the heart's activities: it also makes a thumping noise, for example. But we know the blood pumping activity is the function of the heart precisely because it is this activity that was selected for, while thumping is an accidental effect that plays no functional role.

Such an etiological approach to explaining functions has a long history, dating back to Larry Wright's seminal paper, "Functions" (Wright, 1973), which has since been built upon and refined by a number of philosophers of biology, including Ruth Millikan (1989), Karen Neander (1991), Philip Kitcher (1993), Paul Griffiths (1993) and Peter Godfrey-Smith (1994). I draw chiefly on the notion of function offered by Griffiths and Godfrey-Smith, who build on and refine Millikan's notion of a "proper function" as being that activity that explains a trait's origin and maintenance. The first refinement to this definition is Peter Godfrey-Smith's modern history theory of functions, which renders biological functions as "dispositions or effects a trait has which explain the *recent maintenance* of the trait under natural selection" (my emphasis). This draws a helpful distinction between the origin and the maintenance of a trait. After all, many traits originally were selected for in virtue of one activity, with selection eventually favouring a different activity, which eventually becomes responsible for the "recent maintenance" of the trait. A popular example, cited by Godfrey-Smith, is that feathers likely had the original function of providing insulation, and only later did they acquire the function of facilitating flight. In the remainder of this thesis I will refer to this distinction as being between the *original* function and the *primary* function of a trait, with the former being the activity that explains the origin of a trait in the distant evolutionary past, while the latter refers to the current activity that explains the recent maintenance of the trait by natural selection. Sometimes the original function will continue to be the primary function; sometimes selective forces will favour new activity, making it the primary function; and the transition between primary functions will likely be somewhat messy and vague.

The second refinement I adopt to the definition of function is Paul Griffiths' helpful distinction between the function of artefacts and biological entities. Griffiths points out that a designer can place a kind of selection pressure on the morphology of an artefact, even if the competition between alternative designs is hypothetical rather than actual. This lends even an intentional design process some similarities to an evolutionary process. One can then describe the function of an artefact as being the activity or activities that explain its existence in regard to the intended purpose of the artefact, stressing that the actual form could be shaped by intention, trial and error or even accident.

This raises another important distinction between the *intended* function and the *primary* function of a trait or artefact, with the former being the function articulated by a designer, and the latter being the function that explains the recent maintenance of the trait or artefact, as outlined above. This is important because occasionally a designer will specify an intended function, but in actual fact the primary function will be different, particularly when an accidental function turns out to be the activity that has been selected for. For example, an individual might create a massive network of computer networks with the stated intention of facilitating military communication with redundancy and resilience in the case of an attack on the network. Yet, despite a lack of attacks on that network, it spreads in its use amongst non-military users not because of its resilient characteristics, but because it enables easy and low cost communication and the sharing of cat videos among hundreds of millions of individuals worldwide. In this case, the *intended* function is the facilitation of military communications, yet the *primary* function is something like the facilitation of civilian communications. This distinction is particularly useful when it comes to applying functionalist analysis to cultural traits, particularly as many cultural traits are presumed by their practitioners to hold one function (such as conforming with the will of some deity) while actually performing another function (such as signalling group membership to enhance social cohesion), as I will outline in more detail below.

Another account of function that runs in parallel to the etiological account, but will still prove useful in some contexts in this thesis, was offered by Robert Cummins (1975, 2002). Instead of talking about function etiologically in terms of the effect or activity that was selected for, Cummins talks about functions in the context of the causal roles they play in fulfilling the *overall capacity* of some system. In this mode of analysis, the "Cummins function" of the heart is to pump blood, which contributes to the organism's capacity for respiration, which in turn contributes to the overall capacity for sustaining life and enabling reproduction. The thumping noise the heart makes is not its function simply because the thumping noise does not, itself, contribute to some higher level capacity of the organism, but is a by-product of its function to pump blood. Cummins offers a stand-alone account of function that explains the presence of a trait in reference to some *capacity*, but is agnostic on whether that capacity has been selected for by evolution. Cummins functions can also be used to explain the presence of traits that are harmful to an organism, such as the function of a toxin in causing cell death. However, it is possible to fold Cummins functions into an evolutionary account by stipulating that the overall capacity of an organism is determined by natural selection to be survival and reproduction in its environment, and reflect on the manifold activities that contribute to this overall capacity.

Another way of considering Cummins functions is to draw a further distinction between *primary function* and *secondary function*, whereby the primary function is that property or activity that explains the recent maintenance of some feature or traits (as outlined above), while the secondary function is *any other property or activity* of the feature or trait that is less significant, or non-significant, to the recent maintenance of the feature or trait by natural selection, but which still contributes to some capacity the organism has. Secondary functions can still be shaped by natural selection, although what makes them secondary is that the selection pressures shaping the morphology of the primary function overwhelm the selection pressures shaping the secondary function.

6.1.1: Examples

This distinction between primary and secondary functions can help explain not only why birds have feathers, but why those feathers come in a startling array of colours. If the primary function of feathers is to facilitate flight, how can we explain the highly diverse morphology of feathers without suggesting their colour and pattern comes purely down to chance? In this case, the *secondary* function of feathers can be usefully invoked. For example, the secondary function might be to provide camouflage, or it could be to ward off predators, or it could be to signal to members of the same species that this individual is a potential mate. The selection pressures that act on these functions could contribute to the survival and reproduction of the organism, and could have influenced morphology. However, if the secondary activities came at the expense of the primary – i.e. flight – that individual might suffer a greater selective disadvantage. In this case, facilitating flight would be the *primary* function, and camouflage/warding/signalling would be the *secondary* function.

A similar explanation might be offered for the variation in the shapes, and even prices, of artefacts such as cars. The primary function of a car might be to transport its occupants from point A to point B. Yet even a rudimentary and inexpensive car will prove quite capable of satisfying that function. How, then, can the existence of such diversity of automobiles, particularly exorbitantly expensive luxury models be explained? It might be the case that the primary function of a car is to provide transport, but the secondary function is to signal an individual's success and status, not unlike colourful plumage. If, in a modern society, only the most financially successful individuals can afford to own a luxury car, then the possession of such can serve as a signal of that individual's success and status. Yet, a car that looks impressive but is unable to provide transport is clearly failing in its primary function. Thus transport can remain the primary function with signalling being a secondary function. On the other hand, a car that has been chosen purely for display and never for transportation can be said to have its former primary and secondary functions reversed.

It might even be the case that accidental functions – those activities that are by-products of other functions and serve no selective benefit – can become secondary functions, and possibly even primary functions in time given the right circumstances and selective pressures. These would represent cases of "exaptation," to employ the terminology of Stephen Jay Gould and Elisabeth Vrba (1982). This is where a trait that was formerly not selected for comes to be maintained by natural selection, either by promoting or facilitating the existence of some other adaptive trait, or by becoming a primary trait itself. Thus the primary function of feathers may have been insulation, with secondary functions of enabling gliding/flying and signalling, for example. Yet over time, the selective forces favoured gliding/flying over insulation and signalling, and this change in selective pressure would have had an influence on the morphology of the trait. Following a transitional period, where selective pressures for both functions may have exerted an influence on the trait, eventually the selective pressures for gliding/flight overwhelmed those for insulation, shifting the primary function to gliding/flight and relegating insulation to a secondary function alongside signalling.

In a more human context, consider that one of the most important functions that military helmets perform, besides protecting the wearer from flying shrapnel, is to signal their membership to a particular nation. It is even plausible to imagine circumstances where identification of friend from foe is so difficult in the heat of battle that more lives have been saved by the ability to discriminate a soldier's nationality by their helmet than have been saved by the helmet's resistance to shrapnel. In this case, the secondary function might be elevated in importance to be similar to or even greater than the primary function.

6.1.2: Function in biology

We can now use this functionalist framework to explain the existence and features of some biological and, in the next section, cultural traits. Firstly, a textbook biological trait: the heart. As discussed above, the heart has one main activity: to convulse rhythmically, with the effect of pumping blood around the circulatory system. This rhythmic convulsing also has the effect of producing a thumping sound. The *primary* function of the heart is the pumping of blood, as it is this activity that explains the recent maintenance of the heart in the form it has. The heart is a relatively simple example because, as far as we know, its *original* function is still its *primary* function and it has no significant *secondary* functions. A Cummins functional analysis of the heart would agree that the function of the heart is to pump blood, as it is this activity that contributes to the higher level capacity of respiration, and the ultimate capacity of enabling an organism to survive and reproduce.

Another example, as mentioned above, is feathers. Feathers have a number of properties, one of which is to produce a high level of friction with the air when moved. They also often have bright colours. However the *primary* function of feathers is to facilitate flight, as it is the ability of feathers to trap air that explains their recent maintenance. A Cummins functional analysis would agree that the function of feathers is to contribute to the overall capacity of the organism to fly. As mentioned above, the *original* function of feathers appears to have been insulation, and they may maintain this as a *secondary* function in many species to this date. But feathers are even more complex than hearts, because the primary function alone doesn't explain all the morphological differences among token feathers, namely the bright colours and patterns they exhibit. Feathers also often have a *secondary* function of signalling. The colour and patterns produced by feathers facilitate communication to other organisms, such as by signalling to potential mates or rivals of the same species. If one wants to provide an account of the existence and morphology of a variety of feathers that have the same primary function, one must appeal to the secondary function or functions. It may even be the case that the secondary function of signalling overwhelms the primary function of flight, as is the case with the peacock's tail.

These are relatively rudimentary examples, and the complexities of biology will likely make each individual case terrifically more complex. The interplay between primary and secondary functions, the various selective pressures placed on traits, and the complexities of evolutionary trade-offs, pleiotropy, and linkage disequilibrium will make it unlikely that selection will operate on any single feature of a trait in isolation from other features. For example, selection might favour a particular allele that produces a particular colour plumage, yet that allele might influence other traits as well, which would effectively be favoured by selection – not by virtue of their function but by "piggy backing" on the trait that has been selected for. These complexities notwithstanding, a functional analysis can be made compatible with an evolutionary perspective, and can be applied not only to biological traits but also to artefacts and cultural traits – such as behavioural and moral norms.

6.1.3: Taxonomy of functions

It may be useful to give a short taxonomy of the various definitional terms employed to define functions:

> **Original function:** The original function is that activity that explains the existence of trait.
> **Primary function:** The primary function is that activity that explains the recent maintenance of the trait, either under natural or directed selection.
> **Secondary function:** The secondary functions are those activities that help explain some of the features of a trait by virtue of activities that have undergone selection where that selection is weaker than the selection governing the primary function.
> **Intended function:** The intended function is that goal or activity defined by some designer as being the function of an artefact.

6.2: Cultural functionalism

Now that we have a sophisticated account of function, it can be applied to understanding culture, and ultimately morality. Instead of defining culture in terms of its content, or probing the internal justifications that are offered for the existence of cultural norms or institutions, the functionalist perspective defines culture in terms of the function it plays in steering the behaviour of those who conform to it or how it contributes to serving some broader social goal, such as stability or the smooth operation of a social group, or solving a particular problem endemic to social life.

The idea of viewing culture from a functionalist perspective is well trodden ground in sociological circles. In fact, the likes of Herbert Spencer, Auguste Comte, Talcott Parsons and Emile Durkheim, among others, advocated functionalist perspectives on society and culture well over a century ago. According to Spencer, one could think of society as being analogous to a biological organism – in fact, to Spencer's evolutionarily-charged eye, society was almost a super-level of biological organisation – with various norms, customs, traditions and institutions being "organs" that contribute to its overall "healthy" operation (Spencer, 1883). When those cultural and social components are working correctly and in sympathy, they improve the efficiency of the overall society at large. Parsons had a similar view, whereby individuals come to occupy particular "roles" within society, which are shaped by expectations and norms governing their behaviour (Parsons, 1937). These roles form a complementary network that serve the function of contributing to the overall efficient operation of society – although not without perturbations.

Durkheim continued in the functionalist vein, and was primarily concerned with how societies are structured and how they maintain internal stability over time. He developed a functionalist account along similar biological lines to Spencer to account for these phenomena (Durkheim, 1895; Turner, 1995). According to Durkheim, in order to understand the particular constitution of a social institution or custom, or the reason why certain individuals occupied certain societal roles, one must look at the function that these individuals or roles play in the maintenance of the overall society. Durkheim saw that many social customs and norms – including moral norms – played a role in forging interpersonal and community ties, thus facilitating social and cooperative interaction and contributing to the overall operation of society. Along these lines, he famously posited that one of the functions of religion was to promote community and group cohesion (Haidt & Kesebir, 2010).

Functionalism was a popular thesis in sociology through the late 19th and early 20th century, although it was not without its criticisms. Chief amongst them was accounting for the origins of cultural elements such as norms or institutions. This is because it is implausible to consider that the individuals who were involved in the creation of the norms or institutions were aware of the goal to which they contributed; it is hard to imagine the individuals responsible for the creation or propagation of religious beliefs as intending them solely to promote community sentiment and group cohesion. Simplistic functionalist stories also struggle to account for institutions that are inefficient when it comes to promoting social harmony, or institutions that are even counter-productive to social harmony. Likewise, a moral functionalist account along these classical functionalist lines will have analogous problems to solve when it comes to moral norms that are disruptive to fostering social interaction and cooperation. However, many of these problems can be solved by drawing on the etiological notion of function discussed above in section 6.1, as I will elaborate upon below.

6.2.1: What is culture?

The first step in looking at culture and morality from a functionalist perspective is to ask: what is culture *for*? One way to answer this question is to look at the role culture plays in influencing behaviour. While some other animals may share information between themselves or engage in limited directed learning that could be considered as a form of proto-culture, there is no other creature that sees its life as strongly influenced by cultural factors as *Homo sapiens*. There is also no other animal that has adapted to such a staggering diversity of environments around the world, and which engages in such complex social and cooperative interaction. These two facts are not unrelated.

As Kim Sterelny has pointed out, as recently as five million years ago, the last common ancestor of humans and chimpanzees looked a lot like chimps still do today (Sterelny, 2007, 2012). Yet in the intervening years, chimps have changed relatively little while humans have changed to a staggering degree. One of the main contributors to the profound change in *Homo sapiens* appears to be our complex social nature – effectively, our capacity to produce and spread culture (Byrne & Whiten, 1989; Dunbar, 1992, 2003a; Emery, Clayton, & Frith, 2007; Sterelny, 2012). The ability to share information socially enabled our ancestors to evolve adaptive behaviours in a wide range of environments and climates, and enabled them to adapt to relatively rapidly changing physical environments. However, this capacity also produced a more complex social environment and placed greater strain on our cognitive abilities to effectively navigate that environment.

At a certain point, the social environment also became more significant than the physical environment in terms of its impact on an individual's fitness; an individual who was relatively poorly adapted to exploiting the physical environment (i.e. they were physically weak or uncoordinated or lacked an accurate sense of direction, etc.) but who was highly adept at navigating the social environment generally had greater fitness than someone with the opposite traits. As I will discuss in more detail in chapter 15, the tremendous adaptive benefits of such social interaction put a strong selective pressure on the components of our mind that enabled such social interaction – a notion summed up in the so-called "social intelligence hypothesis" (Emery et al., 2007). This social intelligence was a key capacity that enabled the sharing of complex information about the state of the environment, new means to exploit that environment such as tools (both physical and cognitive), and the creation of behavioural norms that guided the behaviour of others in the group: effectively it enabled the creation of culture (Sterelny, 2003).

Employing the functionalist framework outlined above, one can reckon the function of a particular cultural feature or trait by looking at the activity it performs that explains its recent maintenance in the population. What might this activity be? According to Robert Boyd and Peter Richerson, culture has been – and in many ways still is – instrumental in helping to produce adaptive behaviour (R. Boyd & Richerson, 1985; Richerson & Boyd, 2005). Thus, to the degree that the recent maintenance of many cultural traits is because it has produced adaptive behaviour, then the primary function of culture is to produce adaptive behaviour. Of course, that does not imply that the intended function or any secondary functions of culture need to have anything to do with promoting adaptive behaviour.

An example of this function is the food taboos discussed above in section 6.0. There are many more substances that are harmful to one's health if they are ingested than there are beneficial ones. Many of these substances are readily identifiable thanks to evolved heuristics that can identify cues that indicate their unpalatability, such as a foul odour or unappetising appearance. However, these heuristics are far from foolproof; many poisonous substances also appear entirely edible, or even superficially tasty. As such, augmenting these innate mechanisms and heuristics with culturally transmitted information based on (often costly) trial-and-error learning about which substances can be safely consumed, could significantly improve an individual's welfare, including their reproductive fitness.

As such, many food taboos might be explained as being culturally transmitted behavioural norms that reduce the chances that an individual will ingest something harmful. Such an explanation might account for the Fijian food taboo, because many seafoods around the Fijian isles contain toxins that can be harmful, particularly to infants. Such a food taboo might have proven an effective means of preventing such foods from being consumed by those most at risk from the toxins (R. Boyd, Richerson, & Henrich, 2011; Henrich & Henrich, 2010). Seen from this perspective, cultural traits such as food taboos are a form of technology that serve the function of producing adaptive behaviour of the individual.

However, this does not mean that culture necessarily always produces adaptive behaviour. As Richerson and Boyd have illustrated in their models of cultural evolution, cultural variants (i.e. the various motes of information, customs and norms that are transmitted from individual to individual, etc.) undergo their own form of evolution (R. Boyd &

Richerson, 1985; Richerson & Boyd, 2005). New cultural innovations are introduced into a population through a variety of methods – often through trial-and-error experimentation and occasionally through deliberate innovation – and then undergo a process similar to natural selection, whereby some innovations become more prevalent in the population and others diminish in frequency. Individuals adopt new innovations largely through the process of imitation, with a number of biases that favour the spread of some variants over others, such as a bias towards mimicking either the most popular cultural variations or the variations employed by individuals who are considered to be successful or high status. Innovations can also be shared in a more directed manner through deliberate instruction (Sterelny, 2012). There is also a tendency to favour innovations that are visibly beneficial in the current environment, thus introducing a bias into which innovations spread throughout a population. This process of cultural evolution is able to cumulatively build up an impressive body of knowledge that can enable an individual to enjoy a considerable selective advantage over an individual who is required to navigate the world by their wits alone.

While this process of cultural evolution tends to result in the spread of adaptive cultural variants, it is a somewhat imprecise and error-prone process. The processes of imitation are rarely perfect, and individuals do not always imitate the most adaptive traits. In fact, the biases that encourage the imitation of the most common traits in a population or those traits expressed by visibly successful individuals, can promote the spread of maladaptive traits along with adaptive ones. For example, a successful hunter might exhibit a number of behavioural patterns, such as sharpening his spear before the hunt and performing a ritualistic dance to appease the gods. If he proves to be more successful in the hunt than his peers, his behaviours may come to be imitated. Aspiring hunters observing the successful hunter may well be in ignorance of which behavioural traits are the ones that contribute to a successful hunt, and thus mimic the whole suite, thus encouraging the spread of not only the spear sharpening but the ritualistic dancing as well. Once the ritualistic dancing becomes a common practice, then others may come to mimic it as well. In this way a population may even adopt behavioural traits that are neutral or disadvantageous to their fitness. These can continue to persist even over relatively long stretches of time as long as their cost or negative effects are not so serious as to outweigh their positive effects or the benefits of mimicking other positive traits. For example, mimicking the ritualistic dancing might impose a cost in terms of time and resources, but if those who perform the dance also sharpen their spears, and thus are rewarded with an even greater benefit from their hunting ventures, then the ritualistic dancing can be maintained.

Clearly, though, it would be even more advantageous to only sharpen spears and not bother with the costly ritualistic dancing. Yet, unless the link between hunting performance and sharp spears is actively identified, or even indirectly stumbled upon, then it would take a new spontaneous cultural innovation to erode the dancing and preserve the spear sharpening, thus removing the cost of dancing and retaining the benefit of the sharp spears. This could occur, but until it does, the dancing and sharpening may remain entrenched. As Richerson and Boyd point out, if the cost of evaluating whether a trait is beneficial or not is high enough, then redundant and even maladaptive traits can spread via cultural evolution (op. cit.).

There are also other forces that can co-opt traits such as ritualistic dancing – or food taboos – and maintain them in the population even if they exact a cost. One force draws on the benefits of signalling. For example, a food taboo might be introduced because a particular food source proves to be dangerous, such as the shellfish to pregnant women or pigs in times before hygienic farming and slaughtering practices. Yet, such a taboo might be maintained even after the dangers are removed by other technological innovations, such as improved pre-natal screening or hygienic practices. In these cases maintaining the taboo would appear to impose only a cost: that of removing a potentially fruitful source of nutrition from the diet. However, it might be the case that maintaining the taboo signals membership to a particular cultural group, and that signal is made more reliable purely by virtue of the fact it involves some sacrifice: namely that of removing a potentially fruitful source of nutrition from the diet. Such "costly signalling" is a prevalent phenomenon in biology (Zahavi, 1975; Zahavi & Zahavi, 1997). This costly signal might encourage the observer to trust that the signaller is more likely to conform to their entire set of cultural norms, thus making them a more reliable potential cooperator than an individual from an unknown or differing cultural or normative system. This, in turn, might raise general trust levels between individuals employing the same food taboo, thus facilitating social interaction and mutually beneficial cooperative endeavours. As long as the benefit of the signalling generally outweighs the cost of losing the source of nutrition, then the taboo can be maintained.

Thus, an evolutionarily-informed functionalist story can be told about culture, how it develops and the function that various cultural traits serve. Broadly speaking, culture can be seen as a kind of information technology – a body of information and set of tools including biases, rules and practices – that produced behaviour that tended to advance the relative fitness of individuals who conformed with it compared to those who did not employ those cultural tools. This is not to suggest there is only one possible or optimal culture, nor that culture need be composed of a single set of tools. Rather there are many cultures that have emerged, and countless combinations of cultural variants that can compose a culture.

The cultural innovation and evolutionary process is also a relatively messy affair that can result in many sub-optimal solutions to the problems it is attempting to solve. Furthermore, culture can be "corrupted" by those in power and bent to serve their ends rather than the ends of the individuals conforming to the culture, as I will discuss more in chapter 16. To the degree that one individual or a group has power to shape the culture, there would be an incentive for them to steer the behaviour of others to serve their own interests. I do not think it would take much convincing that such a phenomenon exists, and that this corrupting force, along with the regular counter-surges to eradicate such corruption, are potent features in shaping society, particularly since the advent of agriculture, which further enabled the concentration of power and resources in the hands of a minority. Thus, while culture is a powerful tool for advancing the interests of individuals, like any tool, it can be poorly constructed or misused. However, these facts do not undermine the explanatory power of a functionalist perspective on culture.

6.3: Moral functionalism

Drawing on the above, it is now possible to give an account of morality from a functionalist perspective[8]. As I will discuss in the next two chapters, one of the greatest benefits of social living was not only the ability to propagate information about the local environment and the technologies to exploit it, but also the possibility of coordinated and cooperative activity. Through cooperative endeavour, individuals were able to achieve feats that would have been impossible were they to work alone. Also through cooperation could individuals significantly advance their interests, reproductive and otherwise. However, social living is not without its endemic problems, particularly as cooperation and coordination tend to expose individuals to exploitation by free riders and other forms of defection, not to mention the competition over resources from nearby rival bands, which are also strengthened by internal cooperation.

[8] Moral functionalism as used here is distinguished from the moral functionalism offered by Frank Jackson and Philip Pettit in "Moral Functionalism and Moral Motivation" (1995). Where they draw on the philosophy of mind to characterise functionalism, I draw on biology.

The problems of social living are chiefly constituted by those behaviours that can disrupt social harmony, such as causing harm to others and diminishing their ability to pursue their interests (both biological and psychological[9]) as a part of the group. *Homo sapiens*' ability to overcome these problems and live in ever expanded social groups, thus leveraging the power of cooperation, specialisation and division of labour, has been the crowning achievement of our species in terms of promoting our biological and other interests. As I will argue in more detail in chapter 14, evolution shaped our minds to promote altruistic and prosocial behaviour which helps to solve *some* of the problems of social living and facilitates prosocial and cooperative behaviour. These include emotions like empathy, guilt and righteous outrage, which can serve to foster social bonds, encourage costly altruistic acts and encourage the punishment of those who are disruptive to the social order (Haidt, 2003). However, these evolved psychological mechanisms and heuristics are fickle and error prone. As Philip Kitcher puts it, even with our evolved prosocial and cooperative tendencies, there are still "altruism failures" where self-interest overwhelms our altruistic tendencies ultimately to the detriment of cooperation (Kitcher, 2011).

It was thus the cultural innovation of social and moral norms that enabled our species to provide an improved solution to the problems of social living and thus capitalise on our social and cooperative tendencies. There is even evidence that *Homo sapiens* evolved a capacity specifically suited to the creation and adherence of behavioural norms, and a tendency to punish those who transgress those norms (Sripada & Stich, 2005).

As such, in the broad outside-in sense outlined in chapter 4, I propose to define that morality is a cultural technology that has served (and may still serve) the function of *solving the problems of social living such as altruism failures and facilitating prosocial and cooperative behaviour*. This definition is similar to that proposed by Kitcher, who calls "socially embedded normative guidance... a social technology responding to the problem background confronting our first full human ancestors", specifically the problems caused by altruism failures (op. cit.). My definition is also reminiscent of one given by John Mackie who, after pointing out that humans tend to have "limited sympathies" when it comes to interacting with their fellows, suggested the "function of morality is primarily to counteract this limitation of men's sympathies" (Mackie, 1977). As discussed in chapter 5, moral norms tend to have a particular normative form, where they are typically seen as

[9] I discussed various types of interests in chapter 4.

binding and unconditional behavioural rules that possess "practical clout," and which inspire sanction when transgressed. The perceived importance of the moral domain means that moral norms tend to override norms in other domains, and tend to attract a greater magnitude of punishment when breached.

Thus, the primary function of morality has thus been to advance the biological interests of those who conform to it, including their reproductive fitness. However, clearly the converse is not true: not everything that advances fitness is moral. After all, self-interested and harmful tendencies can also advance fitness. Thus, morality is better defined not in terms of advancing fitness, but as *solving the problems of social living in order to facilitate social and cooperative living*. It is in solving these problems that morality has instrumentally contributed to what we might call a Cummins-style "overall capacity" of advancing fitness. This primary function of morality can be broken down into many secondary sub-functions under a Cummins-style functional analysis, such as: how to coordinate social behaviour; how to resolve disputes that can potentially disrupt the group or harm in-group members; how to divide resources; how to reduce cheating and defection; how to structure hierarchies; and many others.

What about norms that appear to take the form of binding behavioural rules that inspire sanction when transgressed, which are considered moral by their adherents, and yet do not appear to serve the function of solving problems of social living? Examples might include norms that promote harm by encouraging disproportionate retribution, or which punish benign behaviours, or norms that entrench privilege in some minority. Such norms do not serve the function of morality as described above, yet they are still deemed by their adherents to be "moral" because they possess the normative form of moral norms. The existence of norms that work against the primary function of morality, or which do not appear to have anything to do with the problems of social living, can be accounted for via the messy processes of cultural evolution. Errors happen, and innovation is rarely directed with explicit knowledge of what the primary function of morality might be. As such, one would expect there to be many counter-productive or sub-optimal norms innovated just as one would expect creatures to occasionally evolve traits that impose a fitness cost. As long as there are enough cultural variants that do help solve the problems of social living, then the culture can tolerate a load of counter-productive or sub-optimal variants as well.

As such, as a definitional point, it is important to draw a distinction between norms that appear to carry practical clout, and yet do not serve the function of morality, and those norms that do. Thus, a norm that is poorly conceived or inefficient when it comes to serving the function of morality can be considered a "sub-optimal moral norm"; it may still possess the normative form of a moral norm, but it is not very good at satisfying the function of morality. However, a norm that serves the interests of a minority at the expense of the function of morality can be considered a "corrupted moral norm," or as "morally corrupt." Such norms may resemble regular moral norms, they may appear to possess practical clout, and they may be considered moral by their adherents, but they simply do not satisfy the function of morality. This is similar to how a law can be said to be unjust, but it can still exist by writ and be enforced by institutional authorities. It is still a law, just not a good one. A corrupt moral norm is still a moral norm, just not a good one. Such norms will play a role in the moral ecology story in terms of explaining some of the moral diversity that is observed in the world, as will be discussed in chapter 16.

Importantly, this functionalist definition of morality does not necessarily preclude other definitions. Indeed, morality might well be about conforming to the will of a deity, or it could be about promoting the greatest happiness of the greatest number, or it could be about conforming with the duties we have by virtue of being rational agents. Or, at least, perhaps morality *ought* to be about one or more of these alternative renderings. The functionalist definition is silent on these matters. However, the functionalist perspective on morality is very useful when attempting to describe and explain the diversity of moral phenomena, which is the primary goal of this thesis. This enables a thoroughly naturalistic approach more akin to behavioural ecology than traditional moral philosophy, although I expect such a naturalistic approach will have some useful implications on moral philosophy, as I will outline in the final chapter of this thesis.

6.4: Problem background

Once one takes a moral functionalist perspective, the natural question to ask is: what best satisfies the function of morality? As with any functional analysis, one needs to examine what Philip Kitcher calls the "problem background" in order to determine what activities might satisfy it (Kitcher, 2011). For example, the problem background for a toaster consists of the challenges involved in heating bread at a sufficient rate to trigger the browning Maillard reaction without sparking combustion, and doing so in an efficient and economical package. The problem background for morality, on the other hand, is somewhat more complex.

As highlighted by the work of sociologists like Durkheim, and more recently by Jonathan Haidt, it appears that a substantial part of the problem background of morality consists in solving the problems of social living, specifically the problems that emerge when a number of unrelated (biologically) self-interested organisms engage in coordinated and cooperative interaction. Individuals will inevitably see conflicts of interest emerge, and groups will inevitably face challenges in coordinating their activity and maintaining group coherence. Individuals and groups do not only have to face the challenges of internal conflict, but also have to deal with competition from neighbouring groups (Gil-White & Richerson, 2002). The manifold pressures that are exerted on social life are impressive, which only highlights the apparent benefits of overcoming them: if the benefits of social and cooperative life were not so great, then the challenges in facilitating it would make it all but impossible. This might be one reason why there are so few species that successfully pursue cooperative interaction with individuals beyond their immediate kin. Yet, those individuals who were able to successfully engage in rich social and cooperative behaviour, and do so in larger and larger groups, possessed a significant selective advantage over those who lived a more solitary existence.

Thus the problem background of facilitating social living is multifaceted, including issues of encouraging and maintaining social cohesion, signalling group membership, coordinating group activity, managing dominance hierarchies and preventing or resolving conflicts of interest within the group, among others. Morality appears to play a role in solving all of these problems in one way or another, as I will discuss in chapter 16. However, there are two problems of social living in particular that appear to be central concerns of every moral system. The first involves initiating and maintaining cooperation amongst biologically self-interested individuals. The second is in coordinating the activities of those individuals to mutual benefit. Because space is limited even in a doctoral thesis, and because these two problems are especially tractable thanks to game theory, I will focus in particular on the dynamics and complexities of the problems of cooperation and coordination over the next two chapters.

Chapter 7: Cooperative Complexity

> Was a deer to be taken? Every one saw that to succeed he must faithfully stand to his post; but suppose a hare to have slipped by within reach of any one of them, it is not to be doubted that he pursued it without scruple, and when he had seized his prey never reproached himself with having made his companions miss theirs.
> - Jean-Jacques Rousseau

7.0: Friends and enemies

During the heat of the desert war in North Africa in 1941, an unwritten agreement spontaneously emerged between the respective British and German scout forces protecting the far southern frontier of the warzone. During the day, squads of infantry and armoured cars from each side would mount probing attacks on the other, seeking weak points and repelling flanking attacks by the main forces in the north. The fighting over the barren terrain of the North African desert was uncompromising and lethal, with each side jockeying for any advantage they could get. However, at 5.30pm each evening the hostilities promptly ceased and combatants hunkered down where they were or returned to their camps behind the front lines.

No attacks were made overnight. Not a rifle shot or artillery round was fired, until at least daybreak. Even though either side could have potentially mounted a dusk or night-time offensive, thus gaining precious ground in a closely fought conflict, they resisted the temptation with remarkable consistency. Both sides were aware that such night-time offensives can be highly risky, the fighting close and deadly, the cost of guarding against attack wearying on the troops and taxing on supplies. Furthermore, should one side begin mounting such operations, the expectation was that the other would likely soon follow suit, exacting high casualties on both sides. Even more remarkable is the fact that the very existence of the ad hoc agreement meant night-time assaults would likely have taken the other side entirely unawares, as they would have been less likely to be on guard against such an action.

The agreement formed was never made in writing, and never officially sanctioned by the upper echelons of command, but it was respected by the soldiers and officers on the ground on each side, and enjoyed high levels of conformity nonetheless. It was a so-called "gentleman's agreement." There were even some reported cases of unwitting breaches of the accord, with one side accidentally opening fire on the other after 5.30pm. In one such instance, the officers swiftly intervened to cancel the action, apologising profusely to their mortal opponents and sanctioning the troops responsible. The agreement didn't last forever, eventually breaking down when a new battalion took over on one of the sides, but it did last while the original officers who settled into the agreement remained in command and was adhered to with remarkable consistency (von Luck, 1989).

The stakes in the conflict in North Africa during the Second World War were clearly high. There would have been tremendous temptation to exploit the very agreement in order to gain an immediate advantage even though both sides acknowledged that the cost of breaching the agreement could outweigh the benefits over the longer term. Yet cooperation was still able to emerge spontaneously, albeit in a fragile state.

As mentioned in the previous chapter, enabling cooperation between individuals or groups with differing or conflicting interests is one of the most significant problems of social living that moral systems have sought to solve. In this chapter I will explore some of the dynamics of cooperation and coordination with the intention of highlighting some complexities that underscore these core problems of social living. It is in the nuances and complexities of these dynamics that I expect we can find insights into how our moral systems and social psychology have been shaped so as to contribute to moral diversity, as will be discussed in the following chapters.

7.1: The problems of cooperation

In the last chapter I proposed that the chief function of morality has been to help solve the problems of social living in order to facilitate greater prosocial and cooperative behaviour. Many people might agree that their chosen moral system serves this function to some degree, although many would also likely state that this is not what morality is about, nor that it is the primary reason they hold the moral attitudes and norms they do. They might cite a commitment to adhere to the will of a deity, or a belief in moral facts, or perhaps a belief in maximising the utility for the greatest number of people. However, I argued in chapter 6 that one reason that there exist the moral systems we observe around the world today is because they have helped to solve many of the problems of social living. If a moral system failed to serve this function to at least some degree – if it inflamed the problems of social living or was antagonistic towards cooperation – then no matter what its presumed justification, I would suggest it is unlikely to have persisted for long in that form.

The facilitation and the maintenance of cooperation are amongst the core problems that moral systems have had to solve. It is probably not surprising, then, that these two challenges have long been identified by philosophers as problems that moral systems have faced. David Hume, for example, alluded to the problem of cooperation in the face of temptation to free-ride or defect with his famous passage:

> Two neighbours may agree to drain a meadow, which they possess in common; because 'tis easy for them to know each other's mind; and each must perceive, that the immediate consequence of his failing in his part, is, the abandoning the whole project. But 'tis very difficult, and indeed impossible, that a thousand persons shou'd agree in any such action; it being difficult for them to concert so complicated a design, and still more difficult for them to execute it; while each seeks a pretext to free himself of the trouble and expence, and wou'd lay the whole burden on others. (Hume, 1739)

This kind of cooperative interaction has been very effectively modelled in game theory by looking at the Prisoner's dilemma. The related phenomenon of coordination has also been modelled by the Stag Hunt. There are many social interactions that are exemplified by these game theoretical models – not least the draining of meadows and the hunting of stags – and it is in elucidating their complex dynamics that we can better understand some crucial facets of the problem background for morality. It is due to these dynamics, and how moral systems have responded to them, that I suggest we can also find the source of a significant amount of moral diversity in the world. Thus, this chapter will take an important technical diversion into the realm of game theory, focusing on these two problems, which will then lead to a discussion of how moral systems have evolved in response.

7.2: The Prisoner's dilemma

The District Attorney knows that Adam and Eve are gangsters who are guilty of a major crime but is unable to convict them without a confession from one or the other. He orders their arrest and separately offers each the following deal: "If you confess and your accomplice fails to confess, then you go free. If you fail to confess but your accomplice confesses, then you will be convicted and sentenced to the maximum term in jail. If you both confess, then you will both be convicted but the maximum sentence will not be imposed. If neither confesses, then you will be framed on a minor tax evasion charge for which a conviction is certain" (Binmore, 2007).

The Prisoner's dilemma, as illustrated by the vignette above, models a non-zero-sum cooperative interaction between two individuals, or "players." In this interaction, they can each choose to either cooperate with each other or not, with the latter referred to as defecting. Thus, in the context of the above vignette, cooperating consists of refusing to confess to the DA. Defecting, on the other hand, consists of confessing the crime, which not only incriminates themselves but their accomplice as well. If the gangsters both opt to cooperate with each other by refusing to confess to the crime, they both benefit by receiving only a minor conviction. If they both defect by confessing to the crime, they both go to jail, although with a slightly reduced sentence. The trademark Prisoner's dilemma twist occurs because on first glance it seems like both gangsters would benefit by cooperating. However, if Adam expects that Eve will cooperate by refusing to confess, Adam can receive a better outcome for himself by defecting and confessing (effectively "dobbing" or "snitching" on Eve), and he can go scot-free. Eve, anticipating this possibility, and also wanting the best possible outcome, would thus be rational to also defect and confess in order to prevent receiving the maximum sentence. Thus, it seems if both gangsters are rational, they will both defect, and they will both head to the slammer.

In slightly dryer terms, depending on the action taken by each player, they receive a payoff, as shown in **Figure 1** below. For example, if Player A defects and Player B cooperates, then A gets a payoff of 5 and B gets a payoff of 0. The actual unit represented by the payoff can vary. In the above vignette, the unit might be considered freedom from incarceration or fines. In evolutionary biology, the payoff is typically couched in terms of reproductive fitness. In general, the payoff is often couched in terms of utility. **Figure 1** shows one example of a Prisoner's dilemma payoff matrix, giving the relative payoffs for the actions taken by both players.

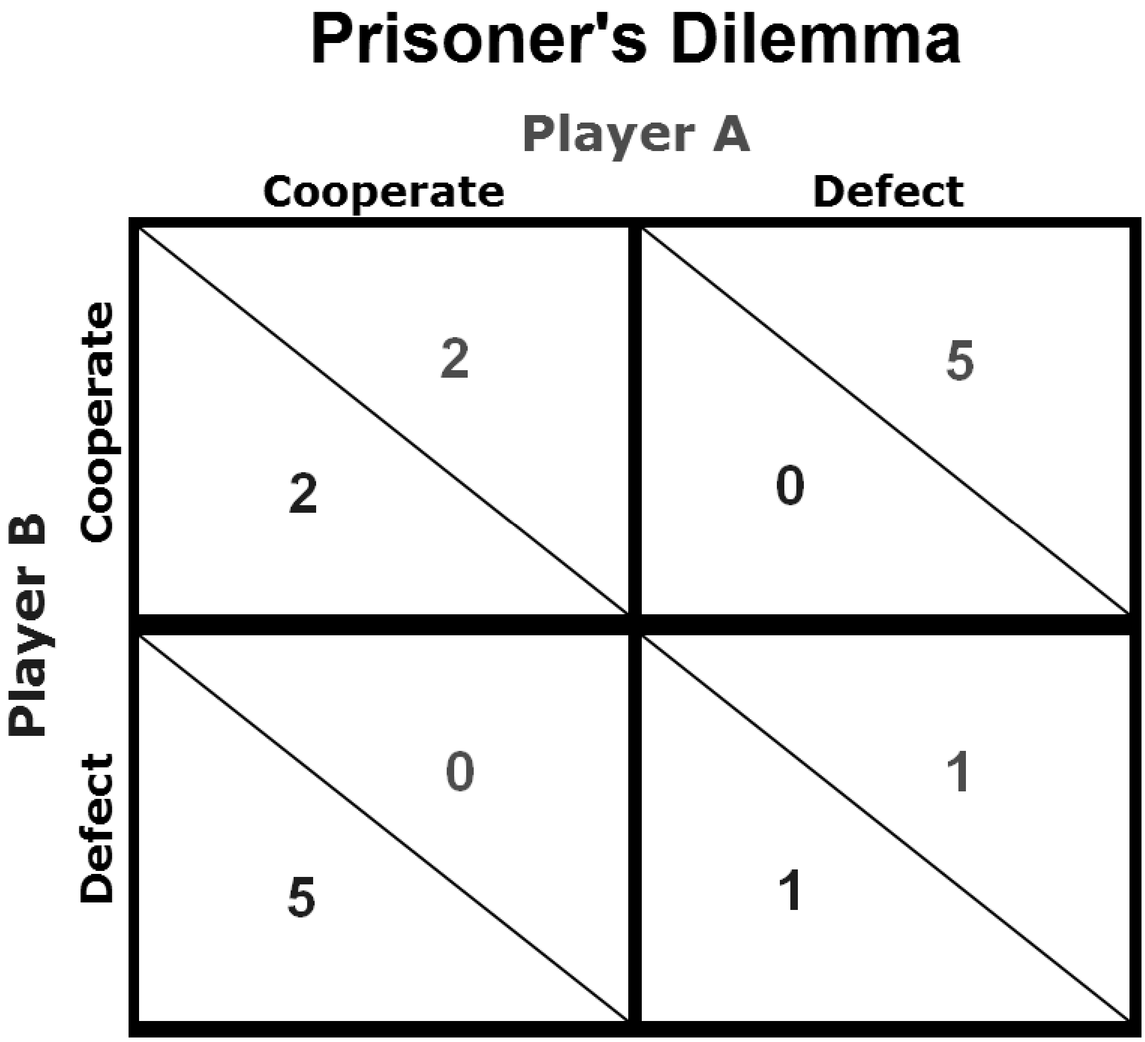
The individual payoffs in the Prisoner's dilemma can be constructed in a variety of ways, but it always follows the general form of T > R > P > S, where:

T = reward for defecting when the other player cooperates, or "Temptation,"

R = "Reward" for mutual cooperation,

P = "Punishment" for mutual defection,

S = reward for cooperating when the other player defects, or the "Sucker's payoff."



	Cooperate	Defect
Cooperate	2 2	5 0
Defect	0 5	1 1

Figure 1: Prisoner's dilemma payoff matrix example.

Thus, if both players cooperate, they both enjoy a reward for their efforts (R). However, if one player defects when the other cooperates, the defector enjoys a greater reward (T) while the cooperator – the "sucker" – receives no reward or, sometimes, a negative cost (S). However, if both defect, they both receive a significantly lower reward than had they cooperated (P). As mentioned above, the curious implication of the Prisoner's dilemma is that rational self-interested individuals will end up mutually defecting and receiving the lowest aggregate reward (PP), whereas they would both benefit by mutually cooperating (RR).

The general form of the Prisoner's dilemma is shown in **Figure 2**.

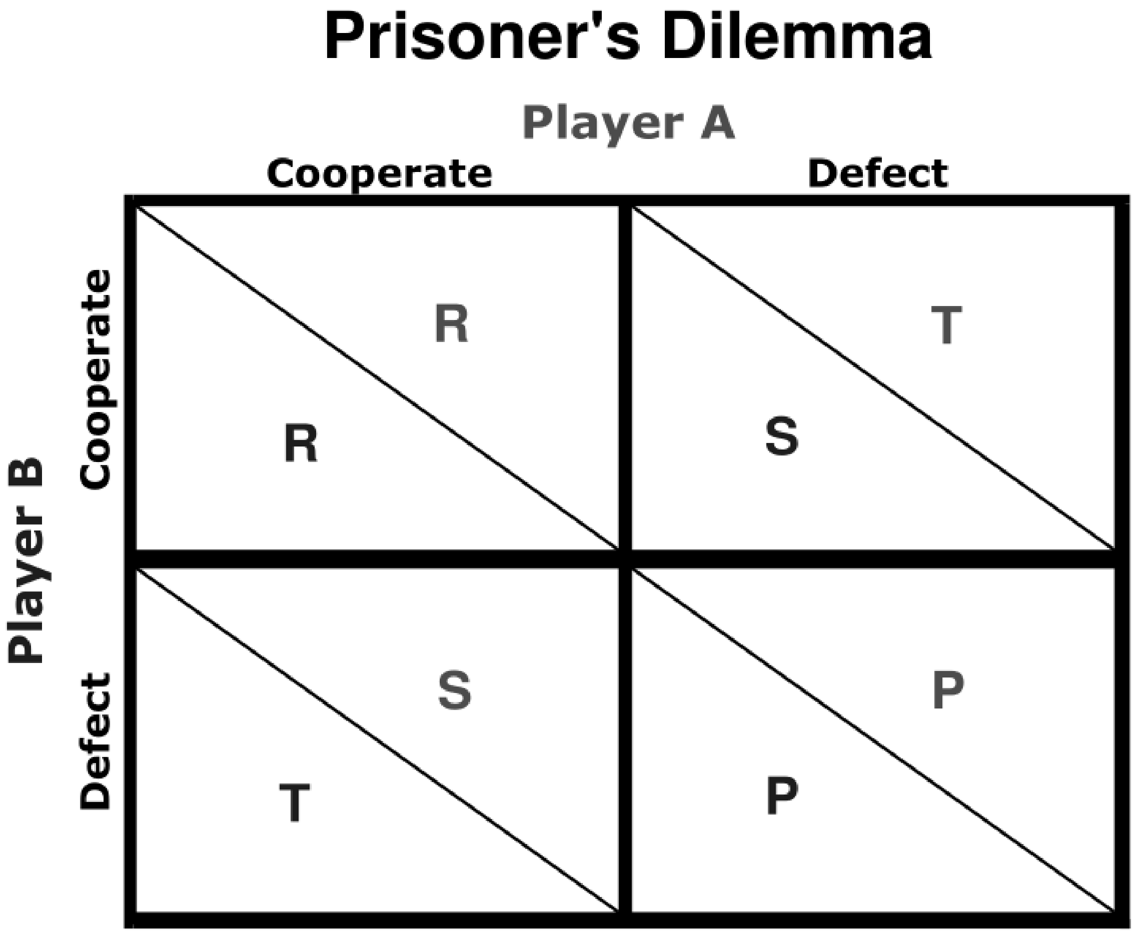


Figure 2: Prisoner's dilemma generalised payoff matrix.

The Prisoner's dilemma models many interactions seen in nature, from the interactions between cleaning fish, *Labroides dimidiatus*, and their "clients", such as the grouper, *Epinephelus striatus* (Trivers, 1971), the warning calls issued by some birds when they spot a predator (Dawkins, 2006), or reciprocal grooming (de Waal, 1982). It also models many interactions in the human sphere, such as between competing companies in setting prices or between superpowers in pursuing an arms race – or an agreement not to fire upon your foes after dusk.

Crucially, it also models a central set of cooperative problems that moral systems typically seek to solve. In many cooperative interactions there is a temptation for one or both parties to defect, thus gaining a potential advantage over their partner. Cooperative hunting is one classic example that featured strongly in our evolutionary past (Sterelny, 2012). If it is the case that hunting large game is physically demanding and dangerous work, then there might be a temptation for an individual in a hunting party to commit the minimal amount of effort and confront the minimal amount of danger if by doing so the party is still likely to be successful in bringing home game. The slack individual can then enjoy the fruits of the hunt without bearing as much of the cost. However, if all members of the hunting party employed the same strategy, there is a higher risk that the hunt would fail, and none of them would enjoy the spoils. Collective defence is another example where a similar dynamic might apply, given an incentive to avoid danger while hoping to reap the benefits of fending off foes.

Another important example is trade, which was one of the key activities that enabled our ancestors to begin specialising their skills, thus reaping the rewards improving the productive output of their groups. While immediate barter might be relatively easy to monitor, if a good is exchanged now in expectation of reciprocation in the future, then there exists an opportunity for the initial recipient to defect and refuse to reciprocate later, thus gaining the benefit of the initial good without paying the cost of reciprocation. If everyone fell for the temptation to defect in such exchanges, it is difficult to see how any meaningful trade could take place, which would likely be to the long term detriment of all the individuals involved.

7.2.1: Nash equilibrium

What is particularly intriguing about the Prisoner's dilemma, and interactions that adhere to its signature form, is that while it looks "rational" for both players to cooperate in order to maximise their overall reward, as soon as one player agrees to cooperate, that individual exposes themself to the possibility of the other player defecting, resulting in the cooperator receiving the Sucker's Payoff (SP). Furthermore, when considering what strategy to employ, Player A must consider what Player B might do. If Player B cooperates, Player A will receive a greater payoff by defecting. Yet if Player B defects, again Player A will receive a greater payoff by defecting rather than cooperating. As a result, defecting always "strictly dominates" cooperating as a strategy, meaning perfectly rational agents who are aware of the payoff matrix, and can anticipate the actions of the individual with whom they are interacting, will always defect.

The conundrum built in to the Prisoner's dilemma is captured by the concept of the Nash equilibrium, named after mathematician John Nash. A Nash equilibrium is a situation in a game where each player cannot improve their outcome by unilaterally changing their strategy, given the strategies employed by the other player or players. For example, if Player A knew that Player B intended to cooperate, then Player A could improve her outcome by defecting, meaning cooperating is not the Nash equilibrium strategy. If Player A knew Player B intended to defect, then Player A could not improve her outcome by cooperating. And this calculus can be performed in a similar way by Player B ruminating over Player A's intentions.

Some interactions modelled by game theory feature many Nash equilibria, but in the Prisoner's dilemma, there is only one: *defect*. It is expected that perfectly rational agents, with full knowledge of the strategies open to them, would employ the strategy that yields the Nash equilibrium. Of course, few if any of us are perfectly rational. However, one need only presume that an agent can remember the payoff they received from previous Prisoner's dilemma interactions, and favour those strategies that yielded the best payoff. In this case, one would expect most agents to gravitate towards *defect*.

This highlights the conundrum inherent in the Prisoner's dilemma: while it is expected that rational agents would choose to always defect, clearly if they both cooperated they would both be better off. In fact, if one looks at the aggregate reward received by both players, it comes from cooperate/cooperate (six points, according to Figure 1), compared with either cooperate/defect (five points) or defect/defect (two points). Cooperate/cooperate is known as the "Pareto optimal" – sometimes called "Pareto efficient" – solution, named after a paper produced in 1906 by Italian economist, Vilfredo Pareto. The Pareto optimal solution is important because it represents the maximum payoff each of the two players can generate in their non-zero-sum interaction.

7.2.2: Iterated Prisoner's dilemma

Things become more complicated, and arguably more interesting, when the Prisoner's dilemma is extended from being a single-shot interaction to a repeated interaction. In the Iterated Prisoner's dilemma (IPD), agents engage in multiple Prisoner's dilemma interactions. The agents can be stipulated to have a memory of past interactions, including the strategies they employed, the payoff they received, as well as memory of the other agents with whom they interacted and the strategies they played. This gives agents an opportunity to adjust their strategies based on experience of past rounds. In some ways the IPD is a more realistic simulation of many real-world cooperative interactions, particularly those between individuals living in the same population, where they are more likely to interact with the same people on multiple occasions. In addition to the condition of T > R > P > S, the IPD also has a condition of T + S < 2R, which prevents cycles of alternating cooperation and defection from being an equilibrium.

If an agent seeks to maximise its payoff over its lifetime, and if that lifetime includes multiple Prisoner's dilemma interactions with other individuals, then it will serve that agent to seek an outcome other than mutual defection. This is particularly the case if the agent is locked in competition with other individuals. Should two or more individuals be able to maintain cooperation over multiple interactions, they will outcompete others who only ever mutually defect. If the payoff is cashed out in terms of fitness, we can see that there would be a selective advantage favouring those individuals who were able to reap the benefits of cooperation. However, there is a twist to this story. If an individual becomes an undiscriminating cooperator, they become vulnerable to defectors, potentially receiving the SP, which is a lower payoff even than mutual defection.

What, then, is the optimal strategy to employ in the IPD? Robert Axelrod famously sought to find out by running a computer tournament, pitting various strategies against each other to see how they compared. A strategy might be something simple, such as *always cooperate* or *always defect*, or it could be more sophisticated, such as *cooperate with a 10 per cent chance following defection* by the opponent[10]. The strategies were run head-to-head in a giant round-robin to see which garnered the most points after a fixed number of iterations. While there is clearly a benefit to cooperating, there is also a risk. One might have expected this risk to preclude the possibility that any cooperative strategy would succeed in the face of defecting strategies. In actuality, the opposite turned out to be the case.

The intriguing result of Axelrod's simulation was that, even in the presence of many defection-heavy strategies, it was an essentially cooperative strategy that emerged triumphant overall. It was called *tit-for-tat* (TFT), and it was the simplest of all strategies submitted (Axelrod, 1984). An agent playing this strategy always cooperates with the other player except in the case that the other player defected in the previous turn, in which case TFT defects as a form of "punishment." If the other player subsequently cooperates, then TFT will cooperate in the subsequent turn. Effectively TFT repeats the opponent's last move in the following turn. The strategy has been likened to the Golden Rule – often stated as "do unto others as you would have them do unto you" – which is a principle that exists in many known moral systems.

7.3: Evolutionary game theory

Axelrod's initial tournament, where different strategies were pitted against each other and their aggregate payoffs compared, is only one way to gauge the success of various strategies in the IPD. Another is to adopt a more evolutionary approach, which he did in a

[10] A note on nomenclature: I will italicise any names of strategies, and employ abbreviations otherwise.

subsequent experiment. Axelrod ran a variation of the original tournament where strategies followed a pseudo-evolutionary process. Each strategy was represented by a virtual genome that represented the actions it would take in certain circumstances. Strategies then underwent an analogue of sexual reproduction and left offspring proportional to their success in the current round. Axelrod also introduced the possibility for mutations in order to introduce new moves into the population of strategies. The result was fascinating: nice strategies again triumphed, led by strategies that closely resembled TFT (Axelrod, 1987). However, new and interesting complexities emerged that will prove important to this thesis.

A crucial lesson from evolutionary game theoretic simulations such as these is that the success or otherwise of a strategy is dependent on the other strategies currently operating in the population; i.e. it depends on the state of the "environment" or the "climate," which includes all the other strategies in the population that it might encounter. As Robert Axelrod and W. D. Hamilton observe, "there is no single best strategy regardless of the behavior of the others in the population" (Axelrod & Hamilton, 1981). Clearly, *always cooperate* does very well in a population of other agents playing *always cooperate*. However, it performs very poorly when confronted by a population of agents playing *always defect*. TFT is particularly interesting because it proved to be "robust," defined as meaning it could "thrive in a variegated environment composed of others using a wide variety of more or less sophisticated strategies" (ibid.).

This idea is also captured in the concept of an "evolutionary stable strategy" (ESS), which was originally articulated by John Maynard Smith and George Price (1973). An ESS is a strategy which, when prevalent in a population, is resistant to invasion by other strategies. Richard Dawkins describes the idea of an ESS thus:

> An evolutionary stable strategy or ESS is defined as a strategy which, if most members of a population adopt it, cannot be bettered by an alternate strategy… Another way of putting it is to say that the best strategy for an individual depends on what the majority of the population are doing. Since the rest of the population consists of individuals, each one trying to maximise his *own* success, the only strategy that persists will be one which, once evolved, cannot be bettered by any deviant individual. (Dawkins, 2006)

The concept of an ESS is similar to that of the Nash equilibrium, except it is defined in terms of a strategy that can resist invasion by new strategies that occur in low frequencies. For example, *always cooperate* is not an ESS in the IPD because it is vulnerable to invasion by even a few examples of *always defect*. However, TFT is an ESS because it performs well against itself, and when prevalent in a population it is resistant to invasion by other strategies.

It is important to note that there need not be only a single ESS, and a population can find evolutionary stability with a mix of strategies that are together resistant to invasion. And while TFT is an ESS, it is not the only ESS in the IPD. TFT can weather invasion by most other strategies, but there are strategies that are also stable in the presence of a high proportion of TFT players, such as *always cooperate*. This is because TFT and *always cooperate* have identical payoffs when they encounter each other; i.e. they always receive the reward for cooperating. This similarity means *always cooperate* can co-exist and increase in frequency within a population of TFT via an analogue of "genetic drift." However, should the proportion of *always cooperate* reach a certain threshold, the population becomes vulnerable to invasion by *always defect*, or another nasty strategy that takes advantage of undiscriminating cooperation. The success of *always defect* sees this strategy increase in frequency in the population, driving payoffs down to the floor of mutual defection.

Another example modelled by Robert Boyd and Jeffrey Lorberbaum sees a slightly "nicer" mutant of TFT creep into a population, *tit-for-two-tats* (TF2T) (R. Boyd & Lorberbaum, 1987). This strategy features many of the characteristics that make TFT so successful: it is "nice," in that it is never the first to defect; it is "provokable," in that it punishes defection with counter-defection; and it is "forgiving," in that it will reinstate cooperation after it has been defected against. However, TF2T is slightly *more* forgiving than vanilla TFT, in that it takes two sequential defections before it is provoked into punishing the defector. TF2T performs as well against TFT as TFT does against itself; i.e. they all engage in continual mutual cooperation. However, should a new mutant enter the population, *suspicious-tit-for-tat* (STFT) – which defects on the first round and then performs identically to vanilla TFT – it can disrupt the stability of the system by entering retaliatory wars (one might call them "feuds") with TFT, as they each defect against each other *ad infinitum*. TF2T, on the other hand, can tolerate STFT's opening defection, and can then enter consistent cooperation with it. The result, interestingly enough, is the demise of the vaunted TFT in this population, to be replaced by a polymorphism of TF2T and STFT. This example reinforces the point made above by Axelrod and Hamilton that "there is no single best strategy regardless of the behavior of the others in the population" when it comes to interactions represented by the Iterated Prisoner's dilemma. Or, as Boyd and Lorberbaum put it,

> When two strategies interact with each other the same way that they do with themselves, their relative fitness depends on their interactions with other strategies. Because neither strategy can be best against every possible third strategy, no pure strategy can resist invasion by any combination of strategies. (ibid.)

In his evolutionary simulations, Axelrod also found that some strategies are able to exploit the cooperation of other strategies and gain a short-term benefit, but they incur a long-term cost as they drive the aggregate level of cooperation in the population down, then exposing themselves to more nasty strategies. Populations can easily pass through grand sweeps of mutual cooperation to mutual defection and back again as strategies adapt to the changing environment.

This also produces some other interesting phenomena, such as there being an apparent trade-off between flexibility and specialisation: a strategy that is adept at exploiting the other strategies in its immediate environment might perform very well over the short term – i.e. that particular strategy might be an evolutionary "attractor," meaning the population will gravitate towards employing that strategy. However, that strategy might perform poorly when the population dynamics change, thus turning it into an evolutionary "repulsor." A flexible strategy might be able to tolerate a wider range of environmental conditions but might have a lower fitness than specialists in a particular environment, thus driving its frequency down, or even to extinction, before it gets a chance to shine (Axelrod, 1997).

TFT has another weakness: noise. This represents errors of action or miscommunications between agents. For example, an agent might intend to cooperate but defects instead, perhaps by firing a few artillery shells after dusk in North Africa in 1941. Such noise could easily break the "contract" of mutual cooperation and see a slide into mutual defection. It took immediate action by the commanders in North Africa to prevent such a slide, including the punishment of their own troops for breaching the agreement, presumably because the commanders had the foresight to predict a precipitous slide into mutual defection and the cost it would impose on their forces. Given how successful TFT is in the IPD, it becomes highly likely that TFT will encounter itself at some stage. Yet given the simplicity of TFT, all it takes is one accidental defection – one "trembling hand" that accidently defects instead of cooperates (Selten, 1975) – to drive TFT into a spiral of mutual defection as each player repeats their opponent's last move. Such an outcome resembles a feud, where neither side is capable of breaking the cycle of mutual defection. In fact, in the presence of noise, TFT yields no better payoff over the long term than does any random strategy (Molander, 1985).

Noise is important particularly when modelling real world interactions, which are likely to be a good deal messier than the rarefied world of idealised agents. The introduction of noise further complicates the IPD, and enables other strategies to outperform TFT. For example, in a noisy system, "generosity" can become more successful. Here generosity means allowing some percentage of the other players' defection to go unpunished with reciprocal defection. Another strategy that performs well in the presence of noise is "contrition," which involves a strategy allowing itself to be punished if it mistakenly defects without defecting in turn, thus preventing a feud. Axelrod and Wu have found that generosity and contrition are both effective at countering noise, although their success still depends on the climate (Wu & Axelrod, 1995).

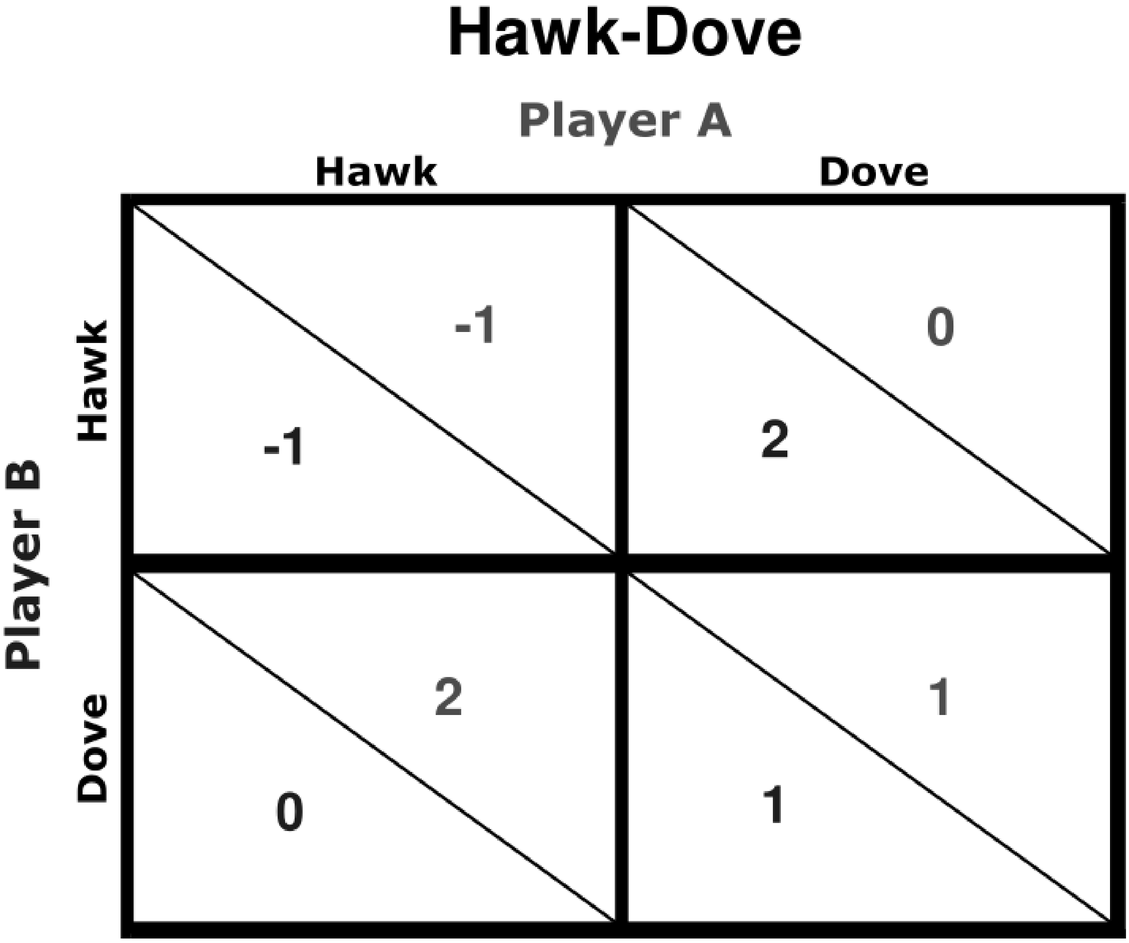
Thus, the success of any one strategy, or a combination of strategies, crucially depends on the environment in which it operates, and that environment includes all the strategies employed by other agents in the population. And as long as new strategies can be introduced via mutation or innovation, then no one strategy can be guaranteed to take over in social interactions that resemble the Iterated Prisoner's dilemma.

7.4: Polymorphisms

Things become even more complex because an ESS need not consist of a single, or "pure" strategy, such as *always cooperate* or TFT, played by the entire population. This can be most simply illustrated using another model from evolutionary game theory, albeit one modelling conflict rather than cooperation: the hawk-dove game, which was championed by John Maynard Smith (1982). The model is described by Richard Dawkins as such:

> Suppose there are only two sorts of fighting strategy in a population of a particular species, named *hawk* and *dove*... Any individual of our hypothetical population is classified as a hawk or a dove. Hawks always fight as hard and as unrestrainedly as they can, retreating only when seriously injured. Doves merely threaten in a dignified conventional way, never hurting anybody. If a hawk fights a dove the dove quickly runs away, and so does not get hurt. If the hawk fights a hawk they go on until one of them is seriously injured or dead. If a dove meets a dove nobody gets hurt; they go on posturing at each other for a long time until one of them backs down. (Dawkins, 2006)

The payoff matrix for the hawk-dove game can be formalised as shown in **Figure 3** below.



Hawk-Dove

		Player A	
		Hawk	Dove
Player B	Hawk	-1 -1	0 2
	Dove	2 0	1 1

Figure 3: Hawk-Dove payoff matrix.

One of the interesting features of this game is that the success of each strategy is "frequency-dependent." This means the payoff a particular strategy receives depends on its frequency in the population. For example, in a population of all hawks, playing dove is highly successful; however, the success of playing dove decreases as the frequency of other doves in the population increases. This is because in a population made up entirely of hawks, a hawk will incur the cost of fighting against another hawk, with them both suffering a payoff of -1. If a dove invades this population, it will retreat whenever it finds a hawk, suffering no penalty. Given the likelihood of a hawk encountering another hawk and suffering the cost of conflict, the dove does relatively well by comparison. Thus, over time, the frequency of doves will be expected to increase, and hawks to decrease. However, if the population is entirely dove, then a hawk can invade, as hawk consistently secures the resource while dove retreats. As such, over time, the population of hawks will increase and doves will decrease. The point of stability in the system occurs at 50 per cent of each strategy. As such, 50/50 hawk/dove is an ESS.

Yet this ESS can manifest in one of two ways. The first is as a polymorphic population playing pure strategies, where half the agents *always* play hawk and the other half *always* play dove, and the other is as a monomorphic, or homogeneous, population playing mixed strategies, where the whole population plays hawk or dove with a probability of 0.5. While the hawk-dove game does not distinguish between these two populations, biological systems that exhibit similar balancing selection tend to be more discriminating (Bergstrom & Godfrey-Smith, 1998). Possibly the paragon example of a polymorphism in biology is the common 50/50 male/female sex ratio found in many species, maintained by negative frequency-dependent selection, which favours the sex with the lowest frequency in the population, thus seeing them gravitate towards a balanced equilibrium (Fisher, 1930).

7.5: Stag Hunting

While the concept of the ESS is incredibly useful at identifying points of stability within a population playing a particular game, this does not mean that a population can necessarily evolve to reach any of those points. A particular ESS might exist in principle and yet it might be impossible for a population that begins with a random mix of strategies to gravitate towards it. This is because populations tend to evolve stepwise in the direction of increasing fitness, avoiding valleys of relatively lower fitness even if it means they cannot reach a more distant fitness peak. For example, an ESS might exist at a maximum fitness peak, yet it might be surrounded by fitness valleys, such that the population will gravitate towards a lesser peak (or multitude of peaks) and settle there rather than traversing the depths of a fitness valley.

Thus, in order for a particular strategy to reach fixation, it needs to possess two features. The first is that it is uninvadable, i.e. it is an ESS. The second is that it is “convergently stable,” meaning it is an evolutionary attractor. A strategy that is convergently stable will tend to outcompete other strategies with which it interacts, causing the strategy to increase in frequency within the population. Strategies that tend to do poorly against other strategies are known as evolutionary repulsors (Axelrod, 1997; McGill & Brown, 2007; Skyrms, 2004).

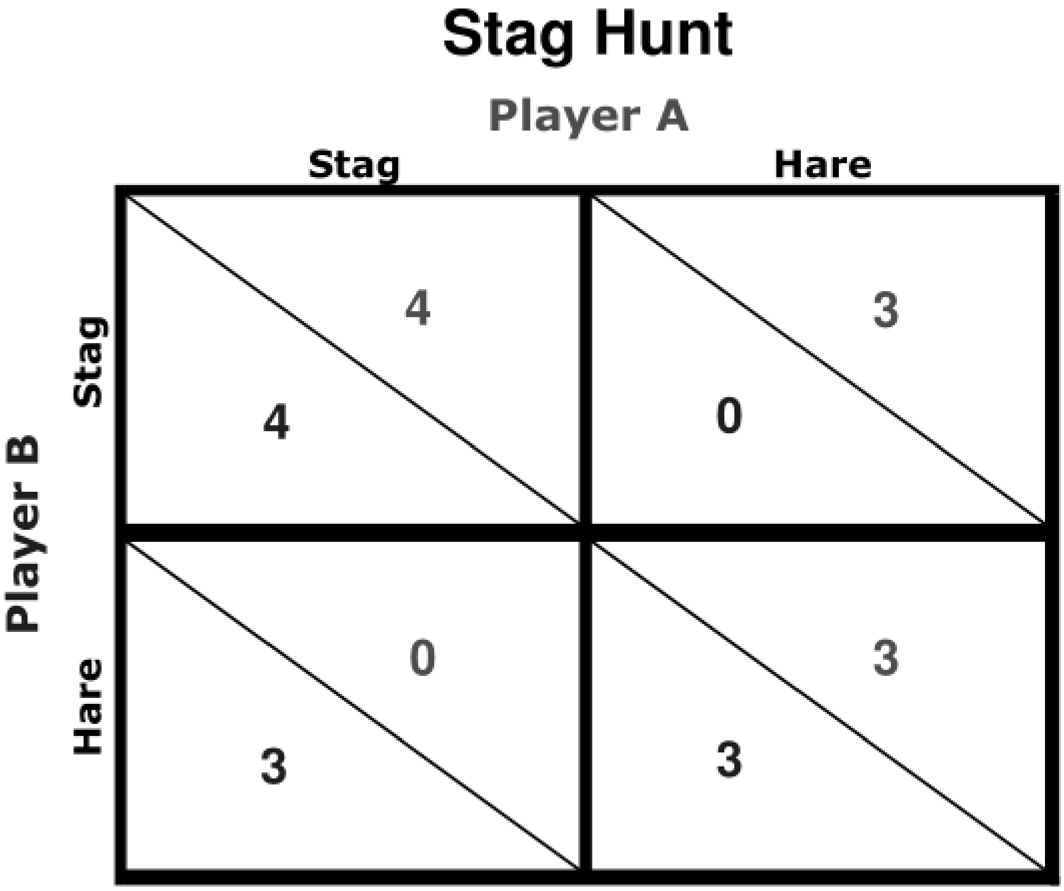
Thus, in order for a population to reach a high fitness ESS, if it can reach it at all, it might have to take a long and meandering path between lower fitness peaks. In the context of facilitating cooperation, this can place limits on what level of cooperation can be stably reached and maintained, particularly if the most cooperative ESS is an evolutionary repulsor, or if reaching it requires traversing fitness valleys.

As I have discussed above, the problem of facilitating and maintaining cooperation is one of the core challenges of social living that moral systems have typically sought to solve. A related problem is coordinating the behaviour of multiple individuals towards a common goal. This problem was alluded to by a near contemporary of David Hume, Jean-Jacques Rousseau, in the quote that tops this chapter. This interaction can be modelled using another game theory construct, the Stag Hunt, named after Rousseau's vignette. Suppose two individuals join together in a hunt for a stag, a prize that would yield more than enough sustenance for the two hunters. However, bringing down a stag is a very difficult task, and it takes both hunters working in concert to have any chance of success. Even then, success is not guaranteed. Suppose, then, that during the hunt one of the hunters is tempted by a nearby hare, the pursuit of which would certainly disrupt any possibility of catching the stag. Even though the hare is a smaller prey – barely enough to provide for a single person – the chance of the solitary hunter catching it is very high. Rousseau wonders whether the hunter might be tempted to abandon the cooperative effort to down the stag in favour of securing a more likely, if materially lesser, prize. Quite presciently, Rousseau thought he might.

Yet, if one hunter did abandon the cooperative endeavour, it would be impossible for the remaining hunter to bring down the stag unassisted. In light of the risk of his hunting partner abandoning him to a hungry fate, the second hunter might be inclined to forego any aspirations for seeking stag and might also go off in pursuit of hare. In this situation, no-one might be inclined to hunt stag, leaving all to go their own way and chase hares alone, even though capturing a single stag would yield a far greater reward for all.

Along with the Prisoner's dilemma, the Stag Hunt, tabulated in **Figure 4**, is a paragon example of a complex dilemma that models coordinated action.

The Stag Hunt is similar to the Prisoner's dilemma, with mutual stag hunting being more productive than mutual hare hunting. However, the Stag Hunt has some different dynamics than the Prisoner's dilemma that introduce new complexities. It models coordination as well as cooperation, with coordinated behaviour (*stag*/*stag* and *hare*/*hare*) being rewarded while uncoordinated behaviour (*stag*/*hare*) receives a lower payoff.



Stag Hunt

		Player A	
		Stag	Hare
Player B	Stag	4 4	3 0
	Hare	0 3	3 3

Figure 3: The Stag Hunt payoff matrix.

As such, the Stag Hunt has two ESSs, one being *hare* and the other *stag*. If a population is settled into the *hare* ESS, it cannot be invaded by a small number of mutants playing *stag*. This is because the hapless *stag* player is likely to interact with a *hare* player, who will promptly pursue their smaller prey, leaving the *stag* with a payoff of 0. However, if the population has settled on *stag*, a solitary *hare* will also do poorly for a similar reason. However, for any one individual, engaging in mutual *stag* hunting is preferable to either mutual *hare* or uncoordinated behaviour. So when the population is sitting at the *hare* ESS, the question becomes: how to tip the population towards *stag*? It turns out that if over three quarters of the population play *stag*, then *stag* will eventually take over and drive *hare* to extinction. However, modelling by Brian Skyrms has shown that if less than three quarters of the population play *stag*, then *hare* will eventually take over (Skyrms, 2004).

Thus, when it comes to evolutionary dynamics, it is not only the ESSs that matter, but also the “basins of attraction” for each individual strategy (or polymorphism of strategies). So two strategies might be ESSs in the strict sense, such that if the entire population plays that strategy then it cannot be invaded by a mutant playing a different strategy. However, it remains for the population to *arrive* at that ESS in the first place; in evolutionary game theory, getting there is often more than half the fun. As the Stag Hunt demonstrates, there may be significant hurdles to overcome in shifting a population from a low payoff equilibrium to a higher payoff equilibrium. In the case of the Stag Hunt, the basin of attraction of *hare* is significantly larger than the basin of attraction for *stag*, meaning most (well-mixed) populations will spiral down to sub-optimal *hare*. Given that *hare* hunting is substantially less lucrative for all involved than *stag*, we might consider the ESS of *hare* hunting to be akin to the Hobbsean State of Nature, in the sense of the starting point from which we strive to break free in order to drive towards more optimal cooperative states, yet acknowledging that there is a fitness valley to be crossed before we can reach that higher cooperative equilibrium.

7.6: Strategic pluralism

Simulations of cooperative interactions such as those mentioned above are common in the game theoretic world, but they do have some limitations when it comes to describing real world phenomena. Many simulations seek to discover the optimal strategy, or mix of strategies, in any particular interaction, as did Axelrod in his original IPD tournament. These simulations can be highly informative, and can reveal unexpected phenomena and dynamics. However, these simulations tend to employ extensive abstractions in order to more easily model the complexities of cooperation, such as assuming that agents are perfectly rational, and that they are optimisers, seeking the best possible payoff in any situation, rather than satisficers, who simply seek a payoff that is "good enough" according to some metric.

In light of this, Bjørn Lomborg undertook a highly sophisticated simulation which sought to explore the dynamics of cooperation in a population of more realistic agents playing the IPD, and which proved highly illuminating in revealing the dynamics of the cooperative systems that emerge (Lomborg, 1996). Lomborg's intention in the study was to answer the question:

> If people are assumed to be selfish, can they, with no central authority, organize themselves to exploit cooperative opportunities by creating norms to regulate interaction? The answer should provide insights into which structures of norms, if any, emerge, how they prevail, and how they break down.

He sought to model the agents in the simulation in a more realistic fashion. As such he sought to avoid modelling them as hyperrational agents with infinite cognitive capacity, and instead modelled them as “boundedly rational” agents. These agents formed their strategies according to economical rules-of-thumb that directed their decisions based on how well they performed in the immediate environment. They also possessed a limited memory of only three previous moves, which enabled them to react to strategies with which they had interacted, but avoided the unrealistic feature of them being able to hone in on particular strategies in order to individually exploit them. Lomborg also simulated the capability of learning by allowing them to imitate successful strategies in their immediate proximity. He also modelled innovation by allowing agents to occasionally randomly change elements of their strategies, thus enabling new strategies to emerge within the population. Finally, he made the system slightly noisy in order to simulate occasional miscommunication or misimplementation of particular moves. As it turned out, noise proved to be a significant feature of his simulations, largely because noise can easily undermine the potency of the mighty TFT, as discussed above.

Lomborg then simulated a population consisting of 2^{20} (1,048,576) agents, which was to ensure that “no individual can have a significant influence on the common environment.” Time was numbered in rounds, which could represent any particular time period in reality, such as days, years or even generations. Within each round, each agent plays two-person IPD games with many (but not all) randomly selected actors from the broader population. While agents do not remember their opponents from previous rounds, they do remember immediately preceding moves within a particular round, thus simulating local interactions.

According to Lomborg: “such modeling assumptions resemble a Hobbesian universe, not of close-knit social relations pervaded by obligations, but of disinterested individuals interacting with abstract ‘others.’” This simulation differed from that of Axelrod’s evolutionary model in that strategies do not themselves replicate. Instead, Lomborg employed a fixed population, but used the capability to innovate and imitate to represent the relative success of particular strategies.

Lomborg measured the performance of each particular agent and strategy, but also monitored the overall cooperation of the entire population by measuring average points per move (APM). The points spread he set was: 5 for Temptation; 3 for Reward; 1 for Punishment; and 0 for the Sucker’s Payoff. Thus, in a population consisting entirely of defectors, the APM would be 1, whereas in a population consisting entirely of cooperators, the APM would be 3, which is the highest APM that can be achieved in the game.

Interestingly, cooperation still reliably emerged even in this more “realistic” simulated world, with boundedly rational agents and with the disruption of noise. In simulations containing low noise (<5%), the APM drove towards the maximum possible level of 3 within 150,000 to 300,000 rounds. However, variation in noise did have a significant impact on the evolution of the worlds. High levels of noise (>5%) proved highly disruptive, particularly to strategies like TFT: “more than 93 percent of the worlds do better than TIT FOR TAT playing against itself under optimal conditions, indicating that *worlds exposed to some degree of noise will generally find much more efficient ways of cooperating than TIT FOR TAT*” (ibid.).

In fact, noise proved to be a highly significant factor in determining both the level of aggregate cooperation achieved, as well as in determining the dynamics of the world. More noisy worlds made downward perturbations more likely, and made it harder for the population to drive cooperation back towards an APM of 3. High noise also increased the amount of variability in the world. Thus noise made it harder to implement a single strategy that could encourage cooperation, prevent defection and guard against noise without sparking off feuds of mutual recrimination. For example, possessing an element of forgiveness made noise less of an issue by making mutual recrimination less likely. However, too much forgiveness can open the door to exploitation by more nasty strategies.

7.6.1: Nucleus and shield

The reason I focus on Lomborg’s study out of the countless other IPD simulations is that the more realistic conditions he implemented resulted in some fascinating dynamics that I believe have bearing on the core problem of social living of how to produce high levels of stable cooperation.

After finding that cooperation can indeed emerge from the populations he simulated, Lomborg then went on to study the dynamics of the worlds populated by these cooperative agents. He found that many populations were stable, not in the sense that they settled into a single ESS that dominated the population, but they did settle into consistently high APMs of over 2.9 for lengthy periods of time, a state Lomborg refers to as “meta-stability.” He found that 60 per cent of the meta-stable populations were “high-yield, stable populations”, meaning they had consistently high APM and they were relatively immune to perturbations, being particularly resistant to invasion by highly defecting strategies. The remaining 40 per cent of stable strategies were only transiently stable, being capable of producing high APM, but being more vulnerable to perturbations from invasion by defectors.

Of particular interest is the fact that the 60 per cent of meta-stable populations employed variations on two particular mixes of strategies. These mixtures consisted of four or five strategies that coexist to produce high APM and high meta-stability. Typical of these mixtures was a combination of highly cooperating, highly forgiving, strategies along with one or more less openly cooperative strategies. The former "nice" strategies would normally be highly vulnerable to defection, and any "nasty" mutant that emerges would be expected to rapidly exploit and overwhelm them. However, when such nasty strategies emerged in the population, their triumph was short lived. This is because most outright nasty and exploitative strategies might perform well against cooperators, but they typically perform less well against themselves. Their own success ultimately becomes their undoing, as many of the agents in the population come to imitate the nasty strategy, thus driving down the APM. At this point, nicer strategies are able to cooperate amongst themselves and outperform the nasty strategies, driving the population back towards adopting the cooperative strategies once more, and returning the population to a high-APM meta-stable state. Lomborg illustrates this phenomenon by tracking the fortunes of one particularly nasty strategy, *69*, once it enters one of the highly meta-stable populations:

> Here I come to the core of the mixed-species solution to the IPD game, the real beauty of cooperation, albeit ever so unintentional. Strategy *69* does very well, but in doing so well, the individuals it exploits the best – those playing *254* ["extremely nice"], *190*, and *178* [both "good-natured"] – give up their own strategies in large numbers and often switch to *69*. In this way, *69* changes the environment to its disadvantage. At the same time, strategies that it does not play well – *236* and *176* – become relatively more attractive, further ruining *69*'s possibilities... Had it been self-sustaining, it might have done alright by using the other actors as stepping stones to dominance, but *69* plays a poor 2.05 APM with itself, still a long way from the minimum population APM of 2.51. (ibid.)

Once *69* was eradicated, the population itself had changed slightly, leaving a mix of strategies similar to the mix prior to *69*'s disruptive emergence, but changed just enough to be immune to the return of *69*.

The rule-of-thumb in a simulation such as this turned out to be that high-performing strategies are almost always exploited by strategies that perform poorly against themselves. And the upshot is that no one strategy alone can produce a high APM while fending off invasion by defectors. The end result is a polymorphism, or "pluralism," of strategies that work together to produce high levels of cooperation and stability: "The point is, that while none of the strategies can hope to cope with most, let alone all, other strategies, *together they might prevail*" (ibid.).

And when one looks at the polymorphism of strategies that are meta-stable, a common pattern emerges, which Lomborg calls the "nucleus and shield". At the core, or the "nucleus," of the population are one or more highly cooperative strategies, which are also highly forgiving, thus providing "unwavering near-total cooperation in the midst of a noisy environment." But co-existing with the nucleus is the "shield," made up of more cautious strategies:

> These strategies are more diverse and, in general, more cautious cooperators. As a group, these strategies generally make fewer points playing themselves than do nucleus strategies playing themselves, but being more cautious, the shield strategies mildly exploit the nucleus strategies. In this way, strategies in the nucleus and shield get the same payoff – thus creating a stable distribution.

Furthermore, the shield effectively protects the highly exploitable nucleus from invasion by nasty mutants. Thus, any nasty strategy that would potentially do well against the nucleus will perform badly against the shield.

The main danger faced by populations such as these is that *too much* cooperation will seep into the system, eroding the shield and making the entire population more vulnerable to invasion:

> Surprisingly, the problem of keeping cooperation going is more a problem of keeping out the "too nice" strategies because they make successful exploitation too easy. Thus, the intermittent breakdown of cooperation is usually caused by too many "too nice" strategies and, perplexingly, this means that cooperation is more stable when there is a solid supply of noxious strategies flushing out the "too nice" strategies. This is seen in the classic predicament of societies going "soft".

7.7: The role of morality

As we can see from the discussion above, it is possible for cooperation to emerge even amongst self-interested individuals. And it can do so spontaneously without any outside interjection, as it did in North Africa in 1941. So if cooperation can occur and be maintained organically, what role is there for morality?

The above discussion hopefully also illustrated how slippery and fragile cooperation is, and how easily disrupted. Sometimes the challenge is to lift a population out of one sub-optimal equilibrium into a more optimal one without falling into a fitness valley en route. Cooperation is also unstable. Even the vaunted *tit-for-tat* can be invaded, eventually driving cooperation down to the floor. Another challenge is in choosing which strategy will yield the highest payoff, especially given that the success of any particular strategy depends on the state of the environment. Yet accurately gauging the state of the population is not necessarily a cheap or easy exercise, and it is one that is very likely to be fraught with error. These errors can also be costly, as "noise" can further disrupt cooperative equilibria. There is often also no one strategy that might be stable on its own, but rather it requires multiple strategies working in concert to lend long term meta-stability, as illustrated by Lomborg.

There is also the terrific challenge of resisting the temptation to defect, particularly in environments filled with cooperators. In fact, the higher the proportion of undiscriminating cooperators in the population, the greater the success for prospective defectors. Yet the defectors themselves pollute the very environment, thus ultimately harming themselves. This phenomenon is akin to the "tragedy of the commons" described by Garret Hardin, whereby self-interest will tend to promote behaviour that diminishes a finite common resource, eventually to the detriment of all (Hardin, 1968). It might seem tempting to simply recommend that everyone always cooperate with everyone else. After all, in such a world, everyone is better off. Yet such a utopian world is ripe for invasion by even a small number of defectors. As long as one cannot prevent new strategies from being innovated (or mutated from existing ones), then the eventual arrival of an invading defector is as good as inevitable. Yet those very invaders risk driving the population to a new equilibrium of mutual defection, which is beneficial to no-one.

Spontaneous cooperation is fragile, and as long as there is a temptation to defect, then it will remain difficult to achieve high levels of cooperation over sustained periods of time. Even the cooperation that emerged between the German and Commonwealth forces in North Africa in 1941 proved fleeting. This is where a system of behavioural norms that are adopted and promoted by members of a population, and backed up by punishment, can be a game changer – and the emergence of normative systems likely was a game changer in our evolutionary past. Kitcher gives one account of how the evolution of language may have prompted early hominins to discuss cases of defection, or “altruism failures”, and formulate new behavioural rules to help prevent them (Kitcher, 2011). One can imagine that the spontaneous example of cooperation that emerged in North Africa in 1941 might have become formalised as a norm, with sanctions against those who transgressed it, if the opposing parties had remained in contact for sufficient time. It is not such a great step to take a behavioural norm that is intended to guide behaviour to prevent altruism failures, and thus advance cooperation, and then to call it a “moral” norm.

It is also probably no accident that *tit-for-tat*, a very successful strategy to lift cooperation in Prisoner’s dilemma-like interactions out of the doldrums of mutual defection, has been formalised as a norm in many moral systems around the world, such as in the Golden Rule. It is probably also no accident that there are many moral norms promoting things like trust, fairness, reciprocity, honesty and other behaviours that facilitate cooperation in interactions resembling the Prisoner’s dilemma.

One crucial element of these moral norms is the spectre of punishment that they carry for those who transgress them. Punishment, which will be discussed in more detail in chapter 9, effectively alters the payoff matrix of cooperative interactions. By adding a cost to defecting, it lowers its payoff to below that of cooperation. Thus, an appropriate level of punishment, if sufficiently enforced, can make it rational for agents to cooperate rather than defect even in interactions that resemble the Prisoner’s dilemma (R. Boyd & Richerson, 1992). As Kitcher points out, when this is the case, then our ancestors did not need to lean on the fallibilities of our psychological altruistic tendencies to promote cooperation, but instead could encourage even Machiavellian egoists – i.e. those who help others purely for self-interested reasons – into the social and cooperative fold (ibid.).

However, our ancestors faced a daunting task in the construction of their systems of moral norms. Which norms are the best at solving altruism failures and at promoting cooperation? The dynamics discussed in this chapter show that arriving at the optimal behavioural strategies is a difficult business, to say the least. A norm that encourages trust can be exploited by behaviour that abuses that trust. A norm that encourages greater suspicion can cause potentially fruitful cooperative endeavours to be missed. The optimal behavioural strategy – or norm that promotes it – will depend on the state of the environment in which it operates. Yet the presence of that behavioural strategy will change the environment itself, much as a new strategy invading a population in the IPD alters the success of the strategies already in that population.

Furthermore, there are few strategies or norms that are likely to perform well in every environment. And sometimes it will be the case that a pluralism of strategies will yield higher levels of cooperation than a single strategy can. These insights are central to the notion of moral ecology. In the next chapters I will explore how moral systems have responded to the complexity of the cooperative landscape, highlighting how these responses have contributed to the existence of moral diversity in the world.

Chapter 8: Moral Ecology

> As it is useful that while mankind are imperfect there should be different opinions, so is it that there should be different experiments of living; that free scope should be given to varieties of character, short of injury to others; and that the worth of different modes of life should be proved practically, when any one thinks fit to try them.
> - John Stewart Mill

8.0: Diversity and ecology

The Galápagos finches – also known as Darwin's finches – are a humble collection of 14 species of small bird endemic to the Galapagos islands that helped to change the way we understand biology (Lack, 1961). While Charles Darwin had the birds collected from the various islands of the archipelago during the second voyage of the *Beagle*, he did not recognise their significance immediately. It was only upon presentation to the famed ornithologist John Gould back in England that he realised them to be new species never before encountered. Yet Darwin observed their startling resemblance to similar finches native to South America, nearly 900 kilometres east across the vast expanse of the Pacific Ocean. He also remarked in *On the Origin of Species* how relatively ill suited they were to the environments on their Galapagos island homes compared to the more fertile climes of South America. If species were supposed to be divinely created with forms already shaped to be in sympathy with their environment – as was the prevailing view at the time – why would these birds "bear so plain a stamp of affinity to those created in America?" (Darwin, 1872).

Yet the finches also appeared to be strikingly different from the birds found on the Cape de Verde archipelagos, which possessed an environment that closely resembled the Galapagos. This apparent disconnect between organism and environment was deeply suggestive to Darwin that creatures migrated and changed slowly over the vast stretches of time, with their present forms offering hints of their ancient origins, rather than being divinely created in their final form.

Somewhat out of place on the Galapagos islands though they were, the finches were not entirely without adaptations of their own (P. R. Grant, 1986). The most famous of these is the shape of their beaks. Some species of Galapagos finch have beaks that are small and pointed, others large and robust, and there is a smooth gradation of forms in between. Yet such diversity between the species is far from being purely stochastic variation. When the finches are considered in the context of their local environments, the shape of their beaks appears to be anything but random. Instead, patterns of correspondence between the beak and various sources of food become apparent. For example, the fine beaks are particularly suited to plucking small insects from the branches of plants, while the larger beaks are adept at cracking the hard outer shell of seeds (Boag & Grant, 1981). Given the function that the beak plays in facilitating the extraction of nutrients from the environment, it is evident that the success of any particular beak shape in executing this function depends on the state of the environment around it: in an environment abundant in seeds and short on insects, a small beak will not serve its host as well as a large beak, for example. And if the environment were to change significantly, one would expect a corresponding change in the selection pressures that might eventually result in the beak changing shape over the course of many generations, and that is precisely what the empirical evidence has shown to be the case (ibid.; B. R. Grant & Grant, 1989).

This particular correspondence between trait and environment is also not a one-way street. As traits change over time, thus potentially better exploiting particular features of their environment, the presence of those traits also influences the environment itself and the other species that inhabit that environment. If, for example, a new beak shape enables a species to very efficiently devour the fruit of its staple plant, there is a chance it might impact the ability of that plant to reproduce and flourish. If that plant then diminishes in availability, or dies out entirely, that would in turn affect the population of birds that feed upon it. As such, species are rarely static for long. Instead they are constantly evolving in response to changes in their environment and, indeed, are constantly influencing the selective environment of their own and other species (Van Valen, 1973). This dynamic interaction between the evolution of species and the evolution of their environment is often referred to as the “Red Queen Hypothesis,” referring to a line in Lewis Carroll’s *Through the Looking-Glass* where the Red Queen informs Alice that in Looking-Glass Land “it takes all the running you can do, to keep in the same place.”

It is not just the “fit” between the trait and the environment that determines how a species evolves. Individuals are not only fighting to survive and reproduce in an environment otherwise indifferent to their fate, but they are also battling other members of their own and other species in the struggle of life. One beak might be rather capable of cracking the husk of seeds, but another beak that is even more optimised for the task will likely give its owner a selective advantage in that environment (P. R. Grant & Grant, 2006).

As Darwin's finches ably demonstrate, in order to understand why a particular trait is the way it is, it is important to not only scrutinise the individual organism in isolation, but to take a broad *ecological* perspective. This is not only the case for the size of a finch's beak, but also for any other heritable trait, even behavioural traits (Birkhead & Monaghan, 2010). Such a perspective brings into focus the many environmental influences – including the influence of the social environment "internal" to the population, as I will discuss below – that shape various traits, and reveals the dynamics of how various traits interact to shape the organism and the environment in turn.

Such an ecological perspective can not only shed light on why species vary *between* environments, but can also help to explain other curious traits, such as the variation *within* a particular species in the same environment. A popular example is the beak size of the large cactus finch, *Geospiza conirostris* (P. R. Grant, 1986). Male large cactus finches have one of two different beak shapes, the first being relatively short and the other being long. As their name suggests, the species feeds on the cactus, prickly pear, although birds with different beaks feed on different parts of the plant. Those with long beaks are able to prise the flesh from the fruit by punching holes through its skin, while those with the shorter beak tear off chunks of the plant and eat the pulp or insects found within. Why, then, does the species have two beak shapes rather than one optimised for its environment? In this case, the success of any particular beak shape depends not only on the state of the physical environment – i.e. the existence of a food source that can be readily exploited by that beak – but also the prevalence of that beak shape within the finch population. Due to the finite supply of food on the finches' home island of Isla Genovesa, if they all had the same shaped beak, they would all be competing for the same food source. As such, as the number of individuals with small beaks in the population increases, the amount of fruit that can be consumed decreases. This, in turn, gives an advantage to those finches that have a large beak and are able to consume the alternative food source. If, however, the number of individuals with large beaks then increases, this gives an advantage to the birds with a small beak. As such, the optimal beak shape varies depending on the relative frequency of that beak in the population. The end result of such a dynamic is a stable polymorphism of beak shapes maintained by negative frequency-dependent selection, whereby there is a certain proportion of small to large beaks that persists across generations. This polymorphism effectively optimises the amount of food available to the population of finches, and more importantly, improves the fitness of individuals within that population (B. R. Grant & Grant, 1989).

In this thesis I take a similar *ecological* perspective on morality. I suggest there are some revealing parallels between the manner in which biological traits vary in response to their environment and how moral systems vary in response to theirs. As discussed in the last chapter, many of the problems that moral systems have attempted to solve, such as how to stably foster cooperative and coordinated activity, are devilishly complex. Thus, to the degree that moral norms will seek to prevent "altruism failures" and promote cooperative behaviour, they will need to respond to the highly complex dynamics of these problems, much in the same way that many organisms have evolved in response to similarly dynamic problems posed by their environment.

In this chapter I will elaborate the metaphorical notion of "moral ecology" and employ it as an explanatory framework that can help to understand how moral systems have themselves evolved in response to the complex problems of social living. In the following chapters I will then look at how systems of moral norms adapt and evolve in response to their environment, and how this process is complicated particularly by the complexity of the social environment. In chapter 13 I will then go on to examine how the dynamics and complexities of social interaction have contributed to shaping our evolved moral psychology, resulting in a diversity of psychological function that also contributes to moral diversity, particularly among individuals within a culture.

8.1: Moral ecology

Social living offers many benefits for individuals, not least enabling cooperative and coordinated behaviour. However, as discussed above, social living is a package deal, bringing with it dangers and pitfalls along with potential boons. As I will elaborate in chapter 14, it appears that evolution has furnished us with a number of psychological faculties that have helped to facilitate social living, not least by encouraging proximate altruistic behaviour (Kitcher, 1998). These include things like "moral" emotions, such as empathy, guilt and righteous outrage (Haidt, 2003), and possibly even an innate moral grammar (Hauser, 2006). While these faculties have helped to solve *some* of the problems of social living and fostered prosocial and cooperative behaviour, many of our evolved psychological heuristics are fickle and error prone. This can result in what Philip Kitcher calls "altruism failures", where self-interest overwhelms our altruistic tendencies, ultimately to the detriment of cooperation and harmonious social life (Kitcher, 2011).

However, we have also evolved another crucial capacity that has helped to foster social and cooperative life: our capacity to create, adhere to and spread culture (Byrne & Whiten, 1989; Dunbar, 1992, 2003a; Emery et al., 2007; Richerson & Boyd, 2005; Sterelny, 2012). One of the things that culture has enabled is the creation of norms that guide our behaviour in a social context. As I discussed in chapter 6, I define morality as cultural technology that has served (and may still serve) the function of *solving the problems of social living such as altruism failures and facilitating prosocial and cooperative behaviour*. In this context, I view moral norms as being a subset of behavioural norms, which can be seen as an expectation to behave in a certain way, with a propensity to attract punishment when behaviour deviates from this norm (Axelrod, 1986; Coleman, 1998). Another way to look at moral norms is as guides that promote certain behavioural strategies in particular contexts, effectively altering the behavioural traits of those who conform to them. Norms can govern almost any sort of behaviour, although the norms of interest in this thesis are those that are generally directed towards solving the problems of social living and promoting prosocial and cooperative behaviour (Haidt & Kesebir, 2010; Sripada, 2005).

Where such norms enable individuals to advance their interests, largely through the benefits of cooperation and coordination, along with lending individuals within groups a competitive edge over individuals in rival groups, we would expect those norms to spread in favour of less effective norms, contingent on the mechanisms of cultural evolution. It is worth stressing again that this process of cultural evolution is prone to error and "corruption," which can produce norms and behaviours that are contrary to the goal of solving the problems of social living and promoting prosocial and cooperative behaviour. Some norms can turn out to be downright destructive to social life, like norms that encourage ritualistic aggression, and others can cause spirals of conflict that harm all involved, such as norms promoting revenge without the possibility of forgiveness. Some norms serve the interests of a minority at the expense of others, such as those that entrench wealth and privilege in a few. However, even in the face of such sub-optimal or corrupted norms, the broad cultural evolutionary forces tend to see strategies that are successful at solving the problems of social living spread throughout a population over time if the benefit they lend is greater than the cost of the harmful norms (Richerson & Boyd, 2005).

Central to the moral ecology picture is the crucial observation from game theory, as discussed in the last chapter, that the success or otherwise of any particular moral norm – or a system of moral norms – in satisfying its function depends on the environment in which it operates. Features of the environment, such as the available resources, the presence of hostile neighbours or the social dynamics of the group, effectively influence the problem background that the moral norm faces. This will in turn influence the relative significance of certain problems, and whether a particular behavioural strategy provides an effective solution to a particular problem. The problem of promoting group cohesion, for example, will have different relative importance to other problems. It will also likely have different optimal solutions depending on things like the distribution of the population over the physical landscape, the presence of competing groups or the social features internal to the population itself.

Likewise in biology, features of the environment influence the problem background for many other traits. A trait that is highly successful at solving the problem of extracting nutrients in one environment – such as a large beak in an environment rich in thick husked seeds – might perform very poorly in another environment – such as a neighbouring island with a different selection of flora. Similarly, internal features of the population – such as the proportion of "hawks" compared to "doves" – can also influence the success of a particular trait. Furthermore, to the extent that many of the problems of social living are reflected in game theoretic models, such as the Iterated Prisoner's Dilemma, it is likely that no single strategy – or norm that promotes that strategy –will perform optimally in *every* environment.

Adding another layer of complexity to this picture is the apparent fact that many behavioural strategies or norms will, in turn, feedback and influence the state of the environment itself – effectively a moral analogue to the process of niche construction (Odling-Smee, Laland, & Feldman, 2003). Not only can the actions of the members of a population change the physical environment, such as by exposing new resources that can be exploited, but their actions can have an even more immediate impact on the social environment. For example, the success of a behavioural strategy, such as trusting strangers, may depend on the behavioural strategies employed by other members of the population, such as whether they are likely to honour or abuse that trust. If that is the case, the very activity of trusting or abusing trust changes the environment such that it alters the payoff of other strategies in the population. Whereas making changes to the physical environment such that it affects the payoff of various behavioural strategies might take many generations of effort, changes can ripple through the social environment relatively rapidly – easily within the span of a single generation, as will be discussed below.

The end result of these various forces is a dynamic melange of interacting moral attitudes, norms and behavioural strategies with various payoffs depending on the environment in which they operate, which in turn remains in a constant state of flux: a kind of dynamic "moral ecosystem," if you will. The upshot of this dynamic is that it encourages the proliferation of some forms of diversity, both in terms of a diversity of norms among cultures that reside in different physical environments, but also a diversity of moral attitudes among individuals even within the same physical environment. The remainder of this chapter will be spent fleshing out the various elements of moral ecology as they contribute primarily to diversity *among* cultures in the moral norms they adopt, and chapter 14 will discuss the impact this moral ecological dynamic has had on our biological – and psychological – evolution, thus helping explain some diversity *within* cultures.

8.2: Environmental dependence

As discussed above, the success or otherwise of a biological trait is largely dependent on the state of the environment in which it exists: a long beak is not going to be as successful for its bearers as a short beak in securing food in an environment devoid of insects, for example. Analogous in moral ecology is the notion that the environment plays a significant role in determining the success or otherwise of a moral norm in terms of solving the problems of social living: a norm that encourages trusting outsiders is not going to be as successful for its bearers as a norm that encourages suspicion of outsiders in an environment where those outsides are disposed to free-ride or defect on that trust. However, the environment is far from being a static feature of the world. In fact, the environment can be wildly complex and dynamic, further complicated by the fact many aspects of it are in constant co-interaction with the traits in question.

So far I have been using the term "environment" in a fairly loose sense, but it is worthwhile tightening its definition somewhat to hone in specifically on those features of the world that are relevant to an organism's or trait's success in that environment. An organism's environment will include many features that have no impact on its fitness, whether it be aspects of the geography, other species with which it never interacts or other features that might change over time without impacting the organism at all. If we wish to pick out only those features of the world that impact the fitness of an organism, we can talk specifically about the "selective environment," which might include things like the relative abundance of nearby resources, the presence of predators or the traits of other individuals of its own species with which it interacts. Likewise, we can talk of the selective environment for a particular trait, referencing only those features of the world that impact the payoff of that trait. Similarly in moral ecology, we can talk of the selective environment as being any features of the world – including social features – that impact the payoff of a moral norm or the behavioural strategy it promotes in terms of satisfying its function. As a matter of convenience, throughout the remainder of this thesis, when I talk about the environment in the biological or moral contexts, I will be referring to the selective environment in particular, unless otherwise noted.

Another term that will be used in a specific sense is "success." In a biological context, the success of a trait in a particular environment is indexed to some problem to be solved. The ultimate problem foisted upon all organisms by the strictures of natural selection is that of survival and reproduction, which can then be broken down into manifold sub-problems into problems such as finding prey or hiding from predators. Success can thus be measured by the degree to which the trait solves that problem more efficiently than competing traits in the population. A similar approach can be taken to evaluate the success of a particular behavioural strategy employed by an individual. Success can be measured by the degree to which that behavioural strategy solves a particular problem efficiently. If one behavioural strategy is superior to an alternative at avoiding predators, for example, then it is more successful. As with biological traits, the ultimate problem is one of survival and reproduction, which can be broken down into more proximate sub-problems, such as finding a mate or deciding when to challenge a competitor or back down.

In the moral context, one must define "success" with a little more care. One wouldn't want to suggest that a successful moral behaviour – or moral norm that promotes it – is only one that advances the reproductive fitness of the individual employing it. Were that is the case, then a norm that encouraged defection in one-shot encounters in Prisoner's dilemma style interactions – such as a norm that permitted cheating strangers whom one is unlikely to ever encounter again – would be deemed morally successful. This is not to say that such behaviours might not be advantageous to a particular individual, at least in the short term. Nor is it to suggest that many individuals might not seek such advantage. Such norms might even exist within some cultures, and they might be deemed successful or morally good within those cultures. However, crucially, such a norm is typically *unsuccessful*, or at least *sub-optimal*, when it comes to fulfilling the ultimate function of morality of solving the problems of social living.

As such, from a biological perspective we can index the success or otherwise of a *behaviour* to the impact it has on advancing the organism's interests, specifically its biological interests. Yet, according to the functionalist definition of morality I have offered in chapter 6, we can index the success or otherwise of a *moral norm* more tightly to how effectively it helps to solve a particular problem of social living. A norm that helps to solve cooperation problems or conflicts of interest within a group, for example, can be considered to be successful, while a norm that causes social disruption or conflict within a group can be considered unsuccessful. Drawing the link between successful behaviours and successful norms is a delicate task, but one that can prove highly illuminating when it comes to understanding morality from an ecological perspective.

This picture of morality is further complicated by acknowledging that organisms are not just in a struggle against the physical environment to survive and reproduce. They are also battling each other. Every environment has a "Malthusian limit," or "carrying capacity," representing the amount of resources available for organisms to exploit (Sayre, 2008). A population can increase in size only to the point it bumps up against the carrying capacity of the environment (Fisher, 1930). Let us say there is an asexual organism that is so apt at survival and reproduction that it leaves behind two offspring before it expires. A population of such organisms would be expected to double in size each generation. Clearly, in an environment with finite resources, there will come a point where the population will exhaust all the resources available to it, and will reach an equilibrium point. At that point, competition becomes a significant factor: it is not enough to simply extract resources from the environment, but to do so in spite of other organisms attempting to do the same. As such, it is important to expand our consideration of the environment to include not only the physical landscape and the resources therein, but also the activities of other individuals – including individuals within the same population – as also being a part of the selective environment. Thus the success of a trait in solving a problem is not only indexed to the environment, but also pegged to other traits in the population that might solve that problem even better.

One useful way to look at the role the environment plays in determining the success or otherwise of traits or moral norms is to see it as posing various *problems* to be solved for the individual on hand. When talking about biological traits, there are many individual problems to be solved, although the ultimate problem posed by natural selection is that of survival and reproduction; i.e. those traits that effectively enable an organism to survive and pass on their genes to future generations are the ones that tend to persist through time. Each of the other adaptive problems to be solved, such as finding prey, avoiding predators, securing a mate etc., ultimately contribute to this overall problem of survival and reproduction. Different environments are clearly going to pose different problems for the organism when it comes to tackling this ultimate challenge.

In an arid environment, for example, the search for water and the problem of retaining it against the desiccating effects of evaporation in hot winds can become major occupations for an organism, with some organisms becoming defined by the traits that constitute their response, such as the remarkable bulge of the baobab tree or the spines on cacti. Meanwhile, securing sufficient water in a wet environment is substantially less onerous challenge, and one that may not pose the same relative adaptive significance compared to other problems that emerge in that environment, such as dodging predators. Some environments might pose entirely novel problems for organisms that live in them – problems that are not found in other environments – such as that of extracting nutrients in the extreme environments surrounding undersea volcanic vents (Van Dover, German, Speer, Parson, & Vrijenhoek, 2002). In addition to the problems posed by the physical environment, there is also the ever-present problem of competing with other members of one's own species, which forms another aspect of an organism's adaptive environment.

Another way to look at the influence of the environment on the success of a trait is in defining the problem background for that trait. The same problem, such as finding water or evading predators, has different dynamics in different environments. Take the textbook example of *Biston betularia*, the famous English peppered moth with two common forms, one a light grey colour and another a mottled dark grey "melanic" form. Given that the moth's normal habitat was populated by light coloured trees with abundant lichen, the light and slightly speckled wings provided excellent camouflage; i.e. the wings proved an excellent solution to the problem of avoiding potential predators. However, over the course of the 20th century the physical environment changed sufficiently in many regions of England, with the bark of trees darkening and lichen dying off due to the emissions of industry. This had the effect of changing the problem background of evading predators for the hapless pepper moth. As a result, the light coloured moths stood out against the blackened trees, making their light pigmentation a poor solution to the problem of evading predators. The relatively rare "melanic" moths thus had a better solution to the problem of evading predators, and thus had a selective disadvantage compared to their mottled brethren, resulting in a change in frequency of the forms within the population over time (Cook, 2003; B. S. Grant, 1999; Kettlewell, 1973). This shows how changes in the problem background can often inspire changes in the traits of an organism – a principle that underlies evolution.

I suggest that an analogous phenomenon occurs when it comes to determining the success or otherwise of moral norms. Where in biology the ultimate problem that traits have to solve is the problem of survival and reproduction, the problems moral norms have to solve are the problems of social living that hamper the possibility of social and cooperative interaction. Different environments – including different social environments – will impact the success or otherwise of moral norms in fulfilling this function. As such, one norm might prove tremendously successful in one environment only to fail miserably in another. For example, a norm that forbids engaging with strangers in potentially costly cooperative ventures might help solve the problem of "nasty" strangers defecting in one-shot interactions. However, such a norm might also end up imposing such a high opportunity cost, in terms of lost cooperative interactions with "nice" strangers, that its costs outweigh its benefits – particularly in environments with a high proportion of "nice" strangers. Like with the peppered moth, one might then expect the less successful norm to eventually be reduced in popularity, ultimately being replaced by the more successful norm.

8.2.1: Environmental dynamics

Some aspects of the environment tend to remain relatively static over the span of multiple generations, such as the geological landscape or the annual rotation of the seasons. Other aspects of the environment are more dynamic, such as seasonal fluctuations, the presence of certain resources like food or water, or the activities of other species, predators or neighbouring competing groups. The degree to which the environment is static or dynamic significantly influences the way that organisms evolve, and the way their traits adapt to their environment (Levins, 1968). Analogously, I hope to show in the following sections that the degree to which the environment is static or dynamic also significantly influences the way that moral norms perform and how they change over time in response to their environment.

One way to describe environmental dynamism in evolutionary terms is as "environmental heterogeneity," or variation over some dimension, whether that be space, time or some other dimension. The challenge for organisms is that a trait that is beneficial in one environment might be deleterious in another, and the organism (or its genes) may not "know" in advance the environment in which it will live. If a feature of the environment is relatively universal and static, then it will tend to shape the traits of an organism through mutation and selection over many generations, honing them for the conditions of that particular environment. However, things get more complex when the environment is more heterogeneous (Meyers & Bull, 2002). There are a number of ways that organisms tend to evolve in response to environmental heterogeneity, such as by developing phenotypic plasticity, or genetic or phenotypic polymorphisms, both of which I will discuss in more detail over the next chapters. However, broadly speaking, one consequence of environmental heterogeneity is diversity, both in terms of genetic and phenotypic diversity (Levins, 1968).

The most heterogeneous environment for many organisms is their social environment, and this is particularly the case for highly social species such as our hominin forebears (Sterelny, 2003, 2012). As will be discussed in detail in chapter 14, not only can an individual's fitness depend greatly on their ability to socially interact with others, but the social environment is prone to change on time scales far shorter than that of other features of the environment. Social interaction is notoriously dynamic and unpredictable, with behavioural responses changing rapidly and often dramatically in response to the behaviour of the individual in question. Some of this complexity is captured in the Iterated Prisoner's Dilemma, as discussed in the previous chapter.

8.3: Internal and external environment

While there are many facets and dimensions to the environment, there are two clusters of features that I would like to focus on that will be useful when it comes to looking at the sources of some forms of moral diversity. The first is what I call the "external environment," which includes those features of the world that are *external* to the population in which the individual in question resides. These features tend to impact the entire population rather than just one or a few individuals within that population. Examples include many aspects of the physical environment, such as the geographical range of the group, climate, presence of available resources, existence of competing groups etc. The key point is that variation in the external environment tends to impact all members of a population or social group in a similar way. So if there is a dearth of rainfall, for example, the problem of finding water is a problem shared by most or all individuals within the population. Likewise, the presence of a nearby group competing for resources impacts most or all members of the group rather than only one or some individuals. As such, individuals within two groups with different external environments might be expected to evolve in different ways in response to their respective selective pressures, thus promoting diversity between groups. Clearly, demarcation of the external environment is fluid and context-dependent, but it typically refers to features of the environment that are external to the population being examined.

The second cluster of features I call the "internal environment," which includes features that are *internal* to the population in which the individual in question resides. These are features of the population itself that impact at least some members within that population. This includes things like the traits or behavioural dispositions possessed by other members of the population. The key point is that variation in the internal environment may not place a selective pressure that nudges the entire population in a particular direction. Rather, variation in the internal environment might be more likely to promote variation *within* the population. For example, consider the variation among species of Darwin's finches living on different islands of the Galapagos archipelago (P. R. Grant & Grant, 2006). Many of the islands have a slightly different environmental profile, with different landscapes and available resources. This external variation has contributed to some of the variation *among* the different species of finch. In contrast, consider the large cactus finch and its characteristic polymorphism in beak size *within* the species. This variation is largely inspired by the internal environment of the population, namely the relative frequency of large to small beaks. The success of an individual cactus finch depends not only on the external features of the environment – i.e. water, nesting sites, abundance of prickly pear cactus trees etc. – but also on the internal environment – i.e. the number of small or large beaks in the population. As such, it is possible to have two individuals with a similar external environment which still exhibit variation because of the nature of their internal environment.

This is not to say that the external environment cannot inspire variation within a population, or that the internal environment cannot inspire the uniform adoption of a particular trait, particularly if it is driven by positive frequency-dependent selection, as I will discuss later. The reason I draw this distinction between internal and external environment is a matter of convenience in clumping together some salient features of the world that tend to influence diversity *among* and *within* populations in different ways. This distinction will be particularly useful in a moral context in explaining some of the sources of variation among groups and some of the variation within groups.

Moral ecology can be summed up as the observation that, to the extent that moral systems serve the function of solving the problems of social living, and to the extent that these problems represent dynamic selective environments, then moral systems would be expected to vary in response to those environments. In the next chapter I will discuss some of the mechanisms by which moral systems do react to their environment through a process analogous to evolutionary adaptation.

Chapter 9: Moral Adaptation

> As many more individuals of each species are born than can possibly survive; and as, consequently, there is a frequently recurring struggle for existence, it follows that any being, if it vary however slightly in any manner profitable to itself, under the complex and sometimes varying conditions of life, will have a better chance of surviving, and thus be *naturally selected*. From the strong principle of inheritance, any selected variety will tend to propagate its new and modified form.
> - Charles Darwin

9.0: Adaptation

One of the key insights offered by Charles Darwin is that where there is heritable variation and selection, there is evolution. As such, we tend to observe in nature organisms changing over time to become better adapted to their environment – both external and internal – as mutation introduces new variations and selection weeds out the less effective at solving the problems of survival and reproduction.

One way to look at the notion of adaptation is that populations of organisms tend to evolve towards a point of evolutionary equilibrium within their environment. This is a point that represents a collection of traits or behavioural strategies that are evolutionarily stable, meaning they are resistant to "invasion" by new traits or behavioural strategies, as discussed in the context of evolutionary game theory in chapter 7. However, this does not suggest that there is only one equilibrium point in any given environment. In fact, as discussed in chapter 7, the dynamics of determining which traits or behavioural strategies are evolutionarily stable are complex indeed, sometimes with multiple equilibria and frequency-dependent interactions complicating matters. Many populations may never even reach a point of stability, but it is useful to think of these equilibrium points as fitness peaks or evolutionary attractors that steer the evolution of traits in their direction.

This concept of an evolutionary equilibrium also does not imply that all members of a population express identical traits, only that the population gravitates to a point where the frequency of many of its traits is not changing significantly from one generation to the next. Indeed, a population can find relative stability with a polymorphism of traits, such as a population of Hawks and Doves. The equilibrium points are also often dynamic, as environmental fluctuations, stochastic forces like genetic drift and the complications of things like pleiotropy – where one gene has multiple effects on phenotype, with some perhaps being beneficial and others deleterious – continually conspire to disrupt equilibria. Yet, as a rule of thumb, populations generally drive towards equilibrium points – sometimes towards multiple points – and the stability at equilibrium is often sufficient that the species tends to persist with a particular constitution for at least as long as the environment remains stable.

9.1: Biological adaptation

A particularly illustrative example of evolution towards equilibrium in action is the case of the pepper moth, *Biston betularia*, mentioned in section 8.2. When the population of pepper moths had a mix of light and dark forms in industrial areas it was in disequilibrium as birds worked to whittle down the number of light-winged moths. As the relative fitness of the dark morph increased, and that trait spread through the population, the moth gravitated towards a new equilibrium. However, the case of *B. betularia* also demonstrates how readily this equilibrium can be disrupted. As industrialisation has decreased over the last few decades of the 20th century, and trees lost some of their sooty pall and lichen began to grow back, the fitness of the light morph again increased, thus nudging the population into disequilibrium (Cook, 2003). This demonstrates the complex dynamics underpinning the interplay between trait and environment that drive evolution.

Another illustrative example of this complex interplay is demonstrated by the bacterium, *Pseudomonas fluorescens*. A classic experiment conducted by Paul Rainey and Michael Travisano (1998) demonstrated how a heterogeneous environment – i.e. one that complicated the problem background for access to resources – disrupted the equilibrium and influenced the evolution of the bacteria into new diverse forms that were in equilibria with their new environments. They populated two sets of beakers with a single strain of ancestral *P. fluorescens* with a "smooth" morph, named such because their surface was relatively smooth. One set of beakers was left undisturbed, thus creating a heterogeneous environment with three distinct "niches," consisting of the body of the growth medium, the medium-air interface and the inside wall of the glass beaker itself. The second set of beakers was shaken, disrupting the niches, effectively producing a homogeneous environment.

Rainey and Travisano found that after a few days three different forms emerged in the heterogeneous environment, consisting of the original "smooth" morph (SM), along with what they called “wrinkly spreaders” (WS) and “fuzzy spreaders” (FS). This is because the unshaken beakers instantiated three different ecological niches, offering three rather different problem backgrounds when it came to extracting nutrients and oxygen from the environment. They found that after several days, SM consistently occupied the body of the medium, while WS formed a self-supporting mat at the medium-air boundary and FS propagated along the base of the beaker. It appeared that each form had traits that were better suited to their particular ecological niches, thus enabling them to out-compete rival forms in those niches. Once the three morphs had become established, they settled into an equilibrium driven by frequency-dependent selection that maintained the three morphs in their respective populations. Meanwhile, in the homogenous environment, only the original smooth morph persisted at equilibrium. This study also underscores that “success” is not only gauged in terms of the calibre of the solution to problems presented by the environment, but also in terms of how that solution compares to competing solutions employed by other individuals in the same population.

Such an evolutionary perspective can be lent to behavioural traits as well as to morphological traits. After all, behaviour also has adaptive significance for organisms; an individual that is unresponsive to the threat presented by predators, or which frightens off potential mates, is unlikely to contribute much in the way of genes to subsequent generations. The fact that behaviour can be an object of selection is an insight that underpins the discipline of behavioural ecology (Birkhead & Monaghan, 2010; Krebs & Davies, 1978). The mechanism of inheritance underscoring behavioural traits might be genetic or it might be another mechanism, but as long as there is heritable variation in behaviour and selection, there will be evolution in behavioural traits. As such, a behavioural strategy that increases fitness would be expected to increase its frequency in the population over multiple generations. This phenomenon has been readily documented in a variety of species (Charnov, 1976; Schoener, 1971; Tinbergen et al., 1962).

Like physical traits, the introduction of new behavioural traits into a population will change the selective environment, particularly the internal environment, as defined in the last chapter, thus changing the equilibrium point or points. Which equilibrium point the population will gravitate towards over time – and it may fail to settle on one at all – depends on the steps by which the population evolves. For example, a behavioural strategy that would potentially have the highest fitness if it were the most common strategy employed within a population might fail to gain a foothold when it is relatively rare (Maynard Smith & Price, 1973). As such, a population might become "stuck," so to speak, at a sub-optimal equilibrium, such as hunting hare in the Stag Hunt (Skyrms, 2001).

9.2: Cultural adaptation

Culture adds another dimension to evolution, although it operates in a somewhat different way to biological evolution. One of the particularly interesting things about *Homo sapiens* is our ability to modify our behavioural traits in a far more fluid and plastic way than can many other creatures (Godfrey-Smith, 1998). Where many animals rely on the multi-generational machinations of genetic mutation and selection to alter their behavioural propensities, we also have the capacity to modify our behaviour via culture, and to do so towards adaptive ends (R. Boyd & Richerson, 1985; Richerson & Boyd, 2005). In fact, our capacity for cultural transmission of information and behavioural tendencies appears to be one of the instrumental factors that has enabled *Homo sapiens* to become the dominant animal on our planet (Sterelny, 2012).

According to Robert Boyd and Peter Richerson, upon whose research I will draw heavily in this section, culture can be considered as a kind of information capable of affecting an individual's behaviour, and which is passed from one individual to another via social transmission. And to the degree that there exists variation – in the form of "cultural variants" in Richerson & Boyd's terminology[11] – that undergo selection, then there will be cultural evolution.

However, cultural variants spread through a population in a somewhat different way than do genetic variants in biological evolution. In the latter, traits are transmitted "vertically" from parents to offspring. If a trait lends an individual a selective advantage relative to individuals possessing other traits, that trait would – all else being equal – be expected to increase in frequency in subsequent generations. Cultural evolution complicates this process somewhat because cultural variants are not just transmitted "vertically" but also "horizontally" from peer to peer. The primary mechanism for cultural transmission is imitation, with a number of factors that influence how and why a particular trait is

[11] I prefer Richerson and Boyd's looser term "cultural variants" to Richard Dawkins' (2006) more popular term "meme," even though they refer to very similar concepts of units of cultural selection. Dawkins' rendering is explicitly intended to parallel genes, in that memes are discrete entities that are passed from individual to individual and which undergo occasional "mutations," and experience selection. However, the parallel might be too strong, as Richerson and Boyd point out. Genes are generally replicated with high accuracy from one generation to the next, with mutations playing a signature role in evolution. Yet cultural transmission is a messy process, with relatively imprecise replication when passed from individual to individual. There is still something interestingly gene-like about cultural variants, but they need not necessarily be thought of as the "elements" of culture, as are memes.

imitated rather than others, such as a bias towards imitating traits that are common in the population or traits that are possessed by visibly successful individuals.

As such, a cultural variant is more likely to spread through a population not only by virtue of its impact on an individual's biological fitness, but also by virtue of its ease of imitation. It does so happen that a cultural variant that has a disastrous impact on an individual's ability to survive and reproduce is less likely to be imitated, but this does not imply that a trait that is likely to be imitated is necessarily going to lend its adherent a fitness benefit. For example, a cultural variant with high "cultural fecundity," i.e. one that encourages an individual to actively teach that same variant to others will, all else being equal, be more likely to be imitated than one that encourages an individual to keep it secret, irrespective of its impact on fitness.

However, culture *does* impact biological fitness, particularly if it makes the production of information about the state of the environment cheaper or more accurate (Rogers, 1988). This makes culture a particularly potent force for adaptation in environments that exhibit a degree of variability or heterogeneity: if the environment is relatively static and/or uniform, then genetic evolution is likely sufficient to enable the spread of beneficial traits through the population; if the environment is *too* variable or heterogeneous, then imitation becomes a less reliable method of acquiring beneficial traits for an individual's particular environmental conditions, thus making cultural evolution a weaker force; however if the variability or heterogeneity are only moderate, then cultural evolution can spread beneficial innovations through a population faster than biological evolution.

Once culture is entrenched, it can also impact fitness purely by virtue of the fact it becomes advantageous for an individual to adopt the cultural practices of their particular group. This introduces a strategic element, in the form of a coordination problem, to culture and the impacts it has on fitness. Language is one example. Because language requires coordination between speakers, it will likely be advantageous (and fitness enhancing) for an individual to adopt the prevailing language of their group no matter what that language is, or how ably it describes the world around them (Sripada, 2007). As Sripada points out, such coordination games often have multiple stable solutions, or as he puts it, they lead to "Multiple Equilibria Strategic Situations", or "MESS's." Sripada also makes the point that "MESSy social domains are a fertile source for the emergence of between-group diversity because when multiple different solutions are available for solving strategic problems, inevitably groups can (and do) reach different solutions to these problems" (ibid.), a point that will be particularly relevant when addressing moral diversity as a result of cultural evolution.

While biological and cultural evolution differ in their processes, they are both sensitive to environmental variation and both enable individuals and populations to adapt to their environments. One of the chief benefits of culture is its speed of operation. Information about the state of the environment, which can contribute to more adaptive behaviours, can spread throughout a population more rapidly via cultural rather than biological evolution. Another way of expressing this is that cultural transmission can effectively reduce environmental translucency, thus enabling individuals to respond more directly to features of their environment that might otherwise have been obscure to them.

As such, like populations of organisms, cultures also tend to gravitate towards a point of adaptive equilibrium with the environment, where the culture is resistant to "invasion" by new cultural variants in the present environment. These cultural equilibrium points work as evolutionary attractors, towards which the population will tend to evolve. However, not all of these equilibrium points will necessarily be desirable, much as there are many sub-optimal equilibria in the Iterated Prisoner's dilemma.

9.2.1: Moral norms

The cultural variants that are of particular interest to this thesis are those that direct behaviour in social situations and contribute to solving the problems of social living – so-called "moral norms." I am particularly interested in how these moral norms evolve and adapt to their environment over time, and in doing so, how they contribute to moral diversity.

While there might be many classes of behavioural norms incorporated into a cultural system, what sets moral norms apart is their perceived binding, overriding and inescapable authority – i.e. Joyce's "practical clout" (Joyce, 2006). However, as discussed in chapters 2 and 3, such practical clout has proven to be problematic to cash out in metaphysical terms from the inside-out perspective on morality. In contrast, from the outside-in perspective taken in this thesis, moral norms are considered to be simply another class of behavioural norms that are innovated and propagated as a part of a broader cultural system, albeit norms that carry a kind of practical clout. There is not necessarily anything particularly metaphysically special about them. They differ from the norms surrounding conventions or games like chess or cricket only in the importance that is typically placed on them, such that they tend to override other norms and concerns, and they tend to attract the stronger punishment when violated. The lack of metaphysical distinction does not mean these culturally instantiated moral norms are not interesting or important, however. On the contrary, given the crucial role moral norms play in regulating social behaviour and protecting individuals from conflicts of interest with others within their group, their principal importance comes as no great surprise.

9.3: Moral adaptation

As mentioned at the opening of this chapter, one of the key insights offered by Charles Darwin is that where there is heritable variation and selection, there is evolution. We readily observe organisms changing over time to become better adapted to their environment – both external and internal – as selection weeds out the traits that are least effective at solving the problems of survival and reproduction. Likewise, we see an analogous phenomenon when it comes to moral norms changing over time to become better adapted to their environment as selection weeds out the norms that are least effective at solving the problems of social living in their particular environment.

However, these norms do not operate in isolation. As I will discuss in more detail below, to the degree that social interaction is "MESSy," as Sripada would put it, such that the success of a particular behavioural strategy depends on the other behavioural strategies at play within a population, then the success of a particular norm will also depend on which other norms are active within a population. Where two or more moral norms promote strategies that tend to cause conflict, or which exacerbate problems of social living, then one might expect one or both of those norms to be susceptible to challenge from a more successful moral norm in that environment. As such, we would expect moral norms to cluster in relatively sympathetic groups, particularly if the activity of that group of norms is effective at solving some of the problems of social living. Thus, just as individual organisms will come to co-exist within a particular environment, forming an ecosystem, analogously individual moral norms will tend to come to co-exist within a particular environment, forming moral systems.

Like biological traits and cultural variants, moral norms tend to evolve towards a point of evolutionary equilibrium within their environment, whereby innovation and selection operate until such time as the norms become somewhat stable in that environment. In this context, the equilibrium point represents a collection of cultural variants or norms that are resistant to invasion by new cultural variants or norms.

This does not suggest that there is only one equilibrium point in any given environment, or that a moral system necessarily reaches a point of equilibrium, only that equilibrium points serve as attractors, steering the evolution of norms in their direction. In fact, due to the strategic and “MESSy” nature of many interpersonal interactions, one would expect there to often be multiple equilibria.

As I will discuss in more detail below, this also does not imply that the equilibrium points in any given environment will be fixed over time. The environment – both external, but particularly internal – can be terrifically dynamic and can be highly sensitive to feedback. Yet I would suggest that moral systems generally evolve towards an equilibrium point, and the stability at equilibrium is often sufficient that the moral system tends to persist with a particular constitution for at least as long as the environment remains relatively stable.

9.3.1: Cooperation, coalition and competition

There are potentially tremendous benefits for an individual who engages in social interaction with others; not least are the fruits of coordinated and cooperative endeavour. In fact, it is likely that one of the key drivers promoting social behaviour was not only the benefits lent by increasing cooperation within groups, but perhaps even more importantly, the importance of coalitions and competition among groups.

Where one powerful individual can exert his or her will over a weaker individual, they have significantly more trouble doing so against a coalition of individuals working together. This, in turn, creates an adaptive pressure favouring larger and larger coalitions that compete against each other (Bowles, 2009; Kitcher, 1998; Sterelny, 2007). However, maintaining a coalition is no easy feat, and requires regulation of behaviour within the coalition to promote coordinated behaviour directed towards a common goal. Furthermore, membership of a coalition does not mean individuals are suddenly motivated purely to behave in the interests of the group; individuals remain individuals, and are prone to pursuing their own psychological and biological interests, even if doing so can be disruptive to the interests of other group members. As such, self-interest and internal competition threaten to disrupt the group as a whole and must be regulated if the coalition is to be sustained. The triumph of morality, as a cultural technology, is that it has helped to solve many of these problems – although often clumsily and not without some serious self-defeating setbacks along the way.

Furthermore, moral systems themselves also compete. Just as in the biological world, competition is one of the key drivers of the evolution of moral systems. Even in environments with abundant resources, populations of successful organisms will tend to increase in size until such time as they hit a Malthusian limit, at which point they are likely to enter into population-limiting competition over resources with individuals in other populations.

There is evidence that our species' not-so-distant past in the late Pleistocene was rife with inter-group conflict, likely inspired by competition over resources or a desire to remove potential competition over resources (Bowles, 2006, 2008, 2009; Sterelny, 2012). Ironically, this inter-group conflict could have been one of the great drivers of intra-group cooperation, the stable promotion of which is one of the primary functions of a moral system. Thus, to the degree that moral systems have influenced the ability of individuals within a particular moral system to more effectively compete with individuals in other groups that employ a different (or no) moral system, then competition between groups has influenced whether those moral systems survived and how they evolved over time. It is worth noting that there is a whiff of group selection in the above story about inter-group competition and the evolution of more cooperative moral systems. However, where group selection is a decidedly problematic notion in a biological evolutionary context (Dawkins, 2006; Maynard Smith, 1964; West, Griffin, & Gardiner, 2007; West, Griffin, & Gardner, 2007), it is substantially less problematic when it comes to a cultural evolutionary context (Henrich, Boyd, & Richerson, 2012; Richerson, Boyd, & Henrich, 2002).

The existence and disposition of competing groups constitute an important part of the external adaptive environment for moral systems. However, this process is complicated in the case of moral evolution by internal tensions within the population employing the moral system, which serve to perturb the way the evolution unfolds. There are thus two broad opposing forces that work in tension to steer the evolution of moral systems, shaping how they change and how they adapt to their environment. The first, mentioned above, is the competitive advantage that individuals in larger coalitions enjoy over individuals in smaller coalitions. However, it is not just size that matters; a well organised coalition with members who effectively coordinate their actions towards inter-group competitive ends will have a further advantage over more fragmented or disorganised groups. Furthermore, individuals in groups that are more cooperative, and thus likely to be more productive, are more effective at out-competing individuals in less cooperative groups – noting that success is not necessarily gauged in terms of biological fitness but in the spread of the cultural variants (Gintis, Bowles, Boyd, & Fehr, 2003; Richerson & Boyd, 2005). As such, when in the presence of other competing groups, there is an upward pressure on group size, group cohesion and internal cooperation, with a corresponding increase in social complexity (Gil-White & Richerson, 2002).

However, the benefit of being in a coalition does not mean that all individuals within that coalition are expected to be genuinely magnanimous, or that there is no internal competition or rivalry over resources. Even if a highly cooperative coalition is able to successfully secure abundant resources, such as through collaborative hunting or via raiding a neighbouring group, there would then emerge the problem of how those resources are to be distributed within the group. There would likely be temptation for powerful (or self-entitled) individuals to exert their will to secure a greater proportion of the resources than others. Thus a microcosm of inter-group conflict can emerge, with coalitions within coalitions vying for power and influence over the group and its spoils. Furthermore, as levels of altruism increase within a group, so too does the reward for free-riding or defecting against relatively undiscriminating trusting cooperators.

One way to look at this phenomenon is that there are multiple evolutionary equilibria in many social environments, at least one being mutual defection. These equilibria serve as attractors, which drag behaviours – and the norms that promote them – towards those points. For example, in a population of highly trusting cooperators, the success of defection will likely result in that behaviour spreading in that population until everyone is at the equilibrium point of mutual defection. Any mechanism that prevents defection would have to enjoy strong enforcement in order to overcome the pull of the mutual defection equilibrium point.

Such free-riding or defection *can* be regulated by mechanisms such as direct or indirect reciprocity, but not without the cost of monitoring reputations (Fehr, 2004), and reputation monitoring gets increasingly difficult as the population size increases (Nowak & Sigmund, 2005). As groups increase in size, so too do the levels of social complexity, along with the challenges of coordinating a large number of unrelated individuals and encouraging them to cooperate without defection taking off (Dunbar, 1998; Sterelny, 2012). As such, there is a second force driving cooperation and group size down, fuelled by internal competition and the temptation to free-ride or defect.

These two forces are in tension with each other, the first driving group sizes and group (and internal environmental) complexity upwards, while the latter works to drag group sizes down to simpler and more easily-managed levels. Yet it is perhaps a testament to the power of inter-group competition that it appears the former force has proven more potent throughout human history (Bowles, 2008), and a testament to the power of moral systems to regulate intra-group conflict to allow groups to grow to the civilisational sizes we see today.

Thus, while there is a powerful impetus to engage in social interaction, such social living poses many serious problems that need to be solved in order for groups to grow larger and more cooperative. One of the elemental problems is coordinating the activity of many individuals such that their endeavour is more beneficial to those individuals than solitary or uncoordinated endeavour – a phenomenon modelled by the Stag Hunt (Skyrms, 2004) and discussed in chapter 7. Another problem is promoting cooperative behaviour where such cooperative interaction raises the prospect of free-riding or defection – a phenomenon modelled by the Prisoner's Dilemma (Axelrod & Hamilton, 1981), also discussed in chapter 7. Other problems include managing clashes of self-interest where such clashes can lead to outbreaks of violence or widespread social disruption, the distribution of resources amongst the group, managing interactions with outsiders, maintaining group cohesion and manifold sub-problems. I will treat some specific examples of problems of social living and some of the solutions that have evolved to solve them in chapter 12. First, I will look at the process of cultural evolution that shapes moral systems and how they adapt to their environment.

9.4: Moral innovation

Like biological evolution, the process of adaptation needs to begin with some variation upon which selection can operate, weeding out the least effective variants over generations. In the biological context variation is introduced chiefly by random variation through mutation. These mutations can be considered to be random innovations that are then tested against the environment for their impact on fitness. Most mutations have a deleterious effect on the fitness of their carriers, but occasionally they give them a selective advantage, thus affording the spread of the innovation through the population over generations.

The process of innovation in cultural evolution is somewhat analogous to biological evolution with a couple of important differences, largely stemming from the fact that biological evolution lacks a designer, whereas culture can benefit from both random and guided variation. Like biological evolution, innovations in the form of new cultural variants can be introduced via what Richerson and Boyd term "cultural mutation," which includes random changes in behaviour or the effects of misremembering some existing cultural variation leading to novel behaviour (Richerson & Boyd, 2005). For example, a pregnant woman might avoid shellfish purely as a matter of taste, and she and her offspring might enjoy a health benefit as a result, and her behaviour might then go on to be mimicked. Or an individual might be taught a relatively inefficient way of fishing only to misremember one step, thus altering it and yielding a greater catch than the original method.

However, one of the hallmarks of *Homo sapiens* is our sophisticated cognitive and problem solving ability, which enables us to analyse problems and propose novel solutions. In contrast to the unguided variation of cultural mutation, Richerson and Boyd refer to this non-random process as “guided variation,” which results from individuals actively experimenting with new behaviours, or tinkering with existing ones, and adopting those that prove more effective. This could be a hunter who experiments with a barbed spearhead, or a forager who fashions a carrying vessel out of leaves to more effectively lug food back to her camp. If such an innovation proves visibly successful, and/or easily demonstrated, it is then prone to be imitated and spread throughout the population, displacing the previous behaviour, according to the mechanisms discussed below.

Such cultural innovations might also concern novel solutions to the problems of social living. Philip Kitcher paints a picture of how groups of early humans might have gathered in the “cool hour” at the end of the day to discuss problems and resolve altruism failures that emerged through their day, and to propose and argue over solutions to those problems (Kitcher, 2011). Such deliberation might have resulted in individuals agreeing to accept some limitations on their behaviour in return for others also accepting similar limitations. New behaviours are endorsed as permissible and others as forbidden, with publicly communicated and enforced norms emerging as a result. Such a “cool hour” story, while plausible, probably overestimates the tendency of humans – Neolithic or modern – to coolly discus conflicts of interest and rationally propose new behaviours or norms that would have effectively solved those conflicts without causing new problems. Furthermore, such discussions – if they occurred – would also likely have been influenced by power structures and hierarchies within the group, giving the more powerful and influential members an opportunity to bend or corrupt the norms in their favour. Such discussions may have been in the context of a trusted authority figure – a high status individual, an elder, a holy person etc – arbitrating over disputes and setting behavioural precedents that then become established as norms. However, there is no guarantee that if such deliberation occurred, it would have resulted in norms that genuinely contributed to solving many of the problems of social living.

Even if one finds Kitcher's "cool hour" story to be plausible, it seems unlikely that many moral norms throughout history were innovated with the ultimate function of morality – i.e. of solving the problems of social living – explicitly in mind. It is a stretch of the imagination, to say the least, to envisage a group of hunter-gatherers 50 millennia ago discussing norms that might maintain high levels of mutual cooperation without being vulnerable to spirals of defection towards the Nash equilibrium of mutual defection precipitated by tit-for-tat retaliation. Certainly, many deliberations might have been inspired by specific instances of conflict or social disruption. However, this cannot account for the full gamut of moral proscriptions, particularly those that indirectly benefit social and cooperative living, such as norms governing group conformity or purity, not to mention those that are justified in terms of conforming with the will of a deity. Even if the aim of solving problems of social living *was* explicitly in mind, it seems unlikely that even a group of wise and magnanimous individuals would have been able to anticipate all the repercussions of the new norms they innovated and could have obviated any possibility of corruption.

Thus, it seems more likely that new norms would have emerged in more spontaneous ways in the form of patterns of behaviour that were enforced by punishment, and which later became established as norms. A recent example might be the spontaneous agreement that emerged in North Africa in 1941, which already possessed the feature of attracting punishment and contrition when breached. As soon as such implicit agreements get made explicit, they start to look at lot like moral norms. They may have also emerged in response to individual instances of highly socially disruptive behaviour, attracting sanction from those around them first, then becoming established as a norm later.

In fact, many moral norms may never have been explicitly articulated, or discussed in terms of being discrete rules. Many moral norms are internalised, and it is only upon reflection that an individual can attempt an articulation of what principle might be behind them. And even this process is prone to error, as Jonathan Haidt and colleagues have demonstrated with the phenomenon of "moral dumbfounding," where subjects who identify a moral transgression find themselves unable to articulate which moral norm has been breached (Haidt et al., 2000). These proto-norms might then be acted upon by the forces of cultural evolution to determine which ones persisted and which faded away.

Other norms may have emerged via random variation of behaviour or existing norms, or emerged in response to issues altogether removed from thoughts of cooperation or social cohesion, and were then maintained and entrenched because they happened to lend some benefit to social existence. For example, many thinkers have suggested that religious belief has been an important mechanism for regulating social behaviour and improving social stability in groups larger than the tribe (Durkheim, 1915; Henrich, 2009; Kitcher, 2011; Norenzayan, 2010; D. S. Wilson, 2002, 2005). Even though religion appears to have served the ultimate function of morality in this respect, it is likely that religious beliefs emerged (i.e. were innovated) with something other than the function of morality in mind.

As such, it seems more plausible that moral innovation occurred more along the lines of innovation by random variation rather than the guided or semi-guided variation proposed by Kitcher. In fact, moral innovation might be more the product of "experiments of living", as mentioned in the quote from John Stewart Mill (1859) at the opening of this chapter, more moral serendipity than moral design.

9.5: Moral evolution

Once new cultural variants are innovated, they are then fodder for the forces of cultural evolution, which influence which variants are maintained and spread through the population, and which fade away. Likewise for moral innovations. As discussed above, cultural evolution functions in a somewhat analogous way to biological evolution, with a key difference being that cultural variants can be transmitted horizontally between peers rather than solely vertically from one generation to the next. Consequently, the chief mechanism of cultural evolution is social transmission between individuals, whether they are biologically related or not. In the model developed by Richerson and Boyd, much of this transmission is "unbiased," in that many cultural variants are absorbed indiscriminately, such as those of language or meal times. This accounts for a great deal of cultural "inertia" – the similarity between the features of a culture from one point in time to the next.

However, not all cultural variants are absorbed indiscriminately. Rather, there are a number of mechanisms that influence which cultural variants are more likely to be imitated, and therefore are more likely to spread from one individual to the next. Unlike biological evolution, where a mutation that increases fitness will be expected to increase its frequency in a population, cultural evolution is not solely contingent on the impact of the cultural variant on biological fitness. Consider a variant, such as an idea or practice, which improves biological fitness but which is so complex and difficult to learn or remember that it is slow to permeate through a population. Compare that to an idea that has little or no impact on biological fitness but which is easy to learn or remember, and which encourages its hosts to teach that idea to others. All else being equal, we would expect the latter cultural variant to spread more readily than the former.

That said, cultural evolution and biological evolution are closely interwoven, particularly because the ability to engage in social learning enables adaptive variants to spread through a population more rapidly than does biological evolution, and this benefit outweighs the cost of maladaptive cultural variants which can also become entrenched in a culture (R. Boyd & Richerson, 1995).

Richerson and Boyd outline two broad forces that influence the evolution of culture and which cultural variants are likely to be imitated and spread (Richerson & Boyd, 2005). The first is "biased transmission," whereby a cultural variant is more or less likely to be imitated based on a number of factors. Firstly, individuals will be exposed to a great number of behaviours, beliefs, attitudes, ideas and other cultural variants over their lifetime, and they exercise an element of choice over which ones they adopt. The second bias is based on the content of the cultural variant itself. Ideas, practices and beliefs that are more easily remembered or shared are more likely to be imitated than those that are difficult to understand or pass on.

One mechanism driving the biased transmission that influences which variants are imitated is the frequency of expression of the cultural variant in the population, such that more common variants are more likely to be imitated – i.e. a tendency towards *conformity* that appears to be rooted in our biological predisposition to prefer common over uncommon variants (Henrich & Boyd, 1998). Such conformity has been shown to be adaptive in cases where imitation reduces the cost of acquiring adaptive information for both imitators and for innovators (Richerson & Boyd, 2005; Rogers, 1988). This can occur when imitation aids individuals in learning by offering them "case studies" upon which to model their behaviour, and enables the accumulation of improvements to a particular practice which can then be shared and developed individually.

Another mechanism that drives biased transmission is a preference for imitating behaviours exhibited by successful individuals, or at least individuals who are *perceived* to be successful. However, success is often difficult to gauge – we may not be able to directly observe or identify what it is that a successful hunter does that brings them more game, or observe or identify what it is that a successful business person does that brings them more sales – so we often evaluate success in terms of indirect signals of that success. And, as mentioned earlier, "costly" signals that are more likely to be honest tend to be respected more than cheap signals that can be easily produced (Zahavi & Zahavi, 1997). As such, proxies including wealth, popularity, status, conspicuous consumption, lavish gift giving or adornment with expensive physical accoutrements indicate an individual who might be doing things “right,” and that they are someone worth imitating. This preference for imitating the successful can still be observed in action to this very day, and may fuel a desire to know about the preferences and habits of the rich and famous in magazine or website format.

However, not all cultural variants are imitated based on the cultural variant’s simplicity, popularity and propensity for it to be employed by successful individuals. Individuals also engage in their own process of innovation, experimentation, adaptation, trial-and-error learning or social learning of cultural variants. People try new things, they seek advice from others, they tinker, they modify existing practices to suit their tastes and they tweak their behaviour based on what works for them. As such, cultural variants that first and foremost *work* at whatever they are trying to achieve tend to have a greater propensity to be adopted, as do those that are easily taught or demonstrated.

All of the above are likely to influence the evolution of moral innovations and aid in their spread through a population. For example, behaviours and norms that are widely adopted within an individual’s peer group have a tendency to be seen as not only typical but also *right*. Many studies have shown that people will lean on their conformist tendencies even when they override their personal – and often strongly held – attitudes about what is right and wrong. Studies have shown that people will modify their attitudes about right and wrong based on their observation of the behaviour of others or on their impressions of what others believe is right or wrong (Altemeyer, 2006; Asch, 1956). Interestingly, this tendency towards conformity appears to be stronger in some individuals compared to others (Altemeyer, 1981; Storm & Wilson, 2009), an intriguing observation the implications of which I will explore in chapter 13. The heuristic that links conformity to morality may also underpin the observed tendency to preference conformity even in the face of acting contrary to one’s moral belief, such that the perceived obligation to conform overwhelms other moral judgements, sometimes leading to horrific atrocities occurring (Gigerenzer, 2008).

The tendency to elevate the moral attitudes of high status individuals above one’s own is also a phenomenon that has been regularly observed. One famous example is the study conducted by Stanley Milgram whereby subjects were induced to administer crippling, and even fatal, electric shocks to others (while unaware the shocks were faked) in the presence of a high status authority figure in the form of a researcher encouraging them to continue applying the shocks (Milgram, 1974). While there have been numerous interpretations of the results of this study, it appears the presence of the high status figure somehow overrode many of the individuals' personal moral judgements. Such a tendency to defer to status and authority in moral matters might also be behind the phenomenon of moral "elevation" or "awe" which is sometimes applied to individuals perceived to be moral paragons (Haidt, 2003).

Finally, guided variation might play a role in the adoption and spread of moral innovations, particularly if employing those moral variations *works* and is seen to do so by those who employ it. If a particular behaviour or attitude benefits the individual employing it by advancing their interests, say by promoting reciprocation or preventing defection or free-riding, then that behaviour or attitude might be more likely to be adopted and imitated by others. Likewise, behaviours or attitudes that are more readily adaptable to different social circumstances, or which are more easily learnt from others, might be more likely to spread through a population.

The mechanisms of biased transmission and guided variation might well contribute to the evolution of moral systems, but they cannot by themselves tell the full story. These mechanisms have a particular use when it comes to explaining how the class of cultural variants that directly benefit the individual employing them spread throughout a population. It begins with a new variant, which lends some kind of advantage to the individual employing it. This advantage is then adopted by others who observe the success of that individual and then others who adopt the variant as it becomes more commonly expressed in the population. Other cultural variants that appear to serve no beneficial function then ride on top of these functionally efficacious variants, as long as the cost they impose is not too high.

However, these mechanisms of cultural evolution have a harder time accounting for cultural variants that directly contravene an individual's desires or interests, such as a variant that dissuades lying or cheating in cases when doing so would benefit the individual – such as in one-shot Prisoner's Dilemma-style interactions. While it might be beneficial for others to have that individual not lie or cheat, and it might be beneficial to that individual in the long term if their entire community is averse to lying or cheating, lying or cheating often yields the highest payoff in many cooperative interactions.

Furthermore, once the population is disposed to cooperate, that population is a fertile target for a new cultural variant that encourages free-riding or defection, the success of which might see it spread through the population via the mechanisms described above. As such, a new mechanism is required in order to account for the spread of many moral norms, particularly those that appear to curb the short term interests of the individual employing it. This mechanism is *punishment*.

9.6: Punishment

It is no accident that one of the hallmarks of moral norms is that they tend to attract punishment when they are flouted or breached (Axelrod, 1986). In fact, punishment – or the expectation thereof – appears to be instrumental in the maintenance of moral norms in culture, and not just punishment of those who transgress but, crucially, also those who fail to punish those who transgress, so-called "moralistic punishment" (R. Boyd, Gintis, Bowles, & Richerson, 2003; Henrich et al., 2006). Punishment is a central piece of the puzzle that explains how moral norms spread throughout a population and eventually become entrenched – and why moral norms are often viewed as behaviours that deserve condemnation and punishment when breached.

One way to look at punishment is that it changes the payoff matrix of games like the Prisoner's Dilemma by placing a cost on defection. Even if the punishment comes in the form of a revenge defection, as happens when *tit-for-tat* is defected against, the loss in the subsequent interaction can be factored into the overall payoff of the two rounds combined. If that cost is sufficiently high, it effectively makes cooperation a more attractive strategy from a rational standpoint, and one that will tend to yield a greater payoff if the punishment is applied consistently.

However, there is a puzzle when it comes to making punishment work because the act of punishment is often costly itself. An agent playing *tit-for-tat* opens itself to counter-retaliation if it punishes another agent for defecting. As a more concrete example, consider an individual interposing themselves on another in order to punish them. One might imagine that individual would put themselves at risk of retaliation by doing so, and thus be dissuaded from engaging in punishing behaviour. As such, one is faced with a tragedy of the commons, whereby the population might benefit from the public good of widespread cooperation but is disinclined to shoulder the cost of punishment in order to foster that public good. Thus, over time, one would expect punishing behaviour to disappear and the public good of widespread cooperation to erode.

As such, the phenomenon of costly or altruistic punishment is something of an evolutionary puzzle – although it is one that has largely been solved, beginning with the models developed by Boyd & Richerson (1992). Building on these, James Fowler has shown how altruistic punishment can emerge (Fowler, 2005) and Boyd and colleagues have shown how cultural group selection can enable altruistic punishment to be maintained in a population, with it being evolutionarily stable when common (R. Boyd et al., 2003). Further evidence has found that humans readily engage in altruistic punishment of defectors, even at considerable cost to themselves (Fehr & Gächter, 2002). Such punishment also appears to enable stable levels of cooperation greater than that yielded by direct and indirect reciprocity alone (Sripada, 2005). As mentioned above, a crucial part of the picture is punishment of not only those who break norms, but also those who refuse to punish those who transgress norms, the so-called process of moralistic punishment.

Thus the mechanisms of cultural evolution, including costly and moralistic punishment, can enable cultural variants that regulate social and cooperative behaviour to spread throughout the population, particularly variants that help to solve some of the key problems of social living, i.e. moral norms. New behaviours and norms can emerge and spread, causing moral systems to effectively evolve over time to produce more social and cooperative behaviour within groups. It is this cultural breakthrough that has effectively enabled *Homo sapiens* to form into ever larger cooperative social groups, driven by the pressure to compete with less cooperative groups.

However, the above praise of moral norms in promoting social and cooperative behaviour, and as maintained particularly by punishment, is not to suggest that the process is not prone to error. Punishment can entrench virtually any behavioural norm – even one that forbids gazing at hedgehogs by the light of the moon – even though such a norm might fail to help solve the problems of social living, and might even produce socially disruptive or maladaptive behaviour (R. Boyd & Richerson, 1992; Richerson & Boyd, 2005).

As such, one might observe two separate cultures of a similar size and with a similar degree of cooperation, and yet they might instantiate quite different systems of moral norms. Given the nature of conformity bias and punishment-enforced norms, they might also regard their own system as the only permissible one and perceive other systems of moral norms as being not only alien but aberrant. Like most processes, cultural evolution can produce both beneficial and detrimental effects, and as long as the benefits outweigh the costs, the process can be maintained. In fact, the ability of the cultural evolutionary story to explain the existence of sub-optimal or costly norms can be seen as a strength of the theory, just like the ability of biological evolution to explain "poorly designed" and sub-optimal traits makes it a more appealing theory than spontaneous divine creation.

In this chapter I have drawn on the tools of ecology and cultural evolution to look at how moral norms might emerge and spread to become entrenched in moral systems, and how they can change over time. In the next two chapters, I will explore some further complexities of this process that have influenced how a diversity of moral systems have emerged throughout the world.

Chapter 10: Adaptive Complexity

> The guiding motto in the life of every natural philosopher should be, 'Seek simplicity and distrust it.'
> - Alfred North Whitehead

10.0: Environmental dependence

So far I have outlined an ecological perspective on morality, one that draws (if metaphorically) on many of the tropes employed by ecology and evolutionary biology. From this perspective, systems of moral norms can be seen as a cultural technology, the chief function of which is to solve many of the problems of social living that disrupt coordinated and cooperative behaviour within groups. Moral norms themselves can be seen as behavioural guides that are innovated via a range of means, one of which may have been random behavioural variation, which was then enforced and spread via costly and moralistic punishment along with the cultural evolutionary forces of guided variation and biased imitation. Those norms that proved more successful at solving the problems of social living, and in doing so promoted cooperation, would have lent their adherents a competitive advantage over individuals living a less social and/or cooperative existence, not least in terms of their relative fitness. Those socially successful individuals, being the carriers of the cultural variants that make up their moral system, then have an opportunity to spread their moral system more widely.

However, the success of any particular norm, or system of norms, crucially depends on the environment in which it operates, with different environments changing the problem background of facilitating social living. In this sense, moral systems tend to evolve towards a point of evolutionary equilibrium within their environment. However, this is complicated by the fact that many environments have more than one point of equilibrium. The upshot of this dynamic is that many moral norms will enjoy different levels of success when it comes to solving the problems of social living in different environments. And, like biological traits, it is likely that few norms will be successful in every environment. It is to this notion of adaptive complexity that I turn in this section.

10.1: Environmental complexity

As discussed in section 8.2, different environments pose different adaptive problems that favour different solutions. This is as true of systems of moral norms as it is of organisms. A physical environment with limited water will pose different adaptive problems for an animal than one with an abundance of water but limited in nesting sites. As such, a trait that is highly beneficial to an organism in one environment might serve as a hindrance in another. Analogously, different environments pose different social problems to be solved. An environment with abundant large prey, which are difficult to bring down by a solo hunter, will present different social challenges to one with an abundance of smaller prey, which are more easily hunted solo. Cooperative hunting of the larger prey might reap greater rewards than solo hunting of the smaller prey, but it places different demands on social organisation and introduces new problems of how to divvy up the spoils of the hunt. As such, a behavioural strategy – or the norm that promotes it – that is highly beneficial to its users in one environment might serve as a hindrance in another.

10.1.1: External environmental complexity

As discussed in section 8.3, it can be useful to draw a distinction between the external and the internal environment, and distinguish the impact each type of environment has on the evolution of traits or moral norms. Variation in either environment can have a significant impact on how well a moral system serves the function of solving the problems of social living, and both environments have different dynamics, even if the boundary between the two environments is somewhat fuzzy.

For example, every population needs a few basic resources, such as food, water and shelter. Thus, the relative abundance of these resources (which are features of the external environment) will influence the success of a system of moral norms by changing the demands placed on individuals to engage in cooperative activity in order to extract those resources. Some environments might favour cooperative hunting as an important source of food rather than foraging, for example. The presence of large prey can effectively raise the opportunity cost of small game hunting and increase the benefit of cooperative hunting, creating the very scenario modelled by the Stag Hunt. Environments with restricted resources might also raise the stakes of defection or free-riding, turning it from having a minimal impact on the welfare of others to possibly becoming a serious or mortal risk, again altering the payoff matrix in coordination and cooperation games.

The existence of neighbouring groups – another feature of the external environment – can also raise the prospect of competition among groups (or, in an alternative rendering, competition among individuals in the respective groups). As discussed in the last chapter, competition over resources might have been one of the chief upward pressures that encouraged the evolution of norms promoting greater levels of coordinated and cooperative behaviour within groups (Bowles, 2008, 2009). Environmental heterogeneity can also have an impact on the evolution of social and moral systems. If, for example, the abundance of big game varies from one year to the next, the costs and benefits of various behavioural strategies – and the norms that promote them – might also change, and may do so unpredictably, thus promoting more behavioural or normative flexibility or conditional strategies.

10.1.2: Internal environmental complexity

Variation in the internal environment would also be expected to have an impact on which behaviours and norms are more successful at solving the problems of social living. One example of an internal environmental variable is simply group size. The problems of coordination would likely be significantly more complex in larger compared to smaller groups. As group size increases, it also becomes more difficult to monitor the reputations of, and relations between, other group members, eventually reaching a level where internally representing all of the groups' relations becomes prohibitively taxing in terms of time and cognitive resources (Dunbar, 1998; Sterelny, 2007).

Another example of an internal environmental feature that can affect the success or otherwise of a behavioural strategy or norm is the level of trust that exists between individuals in the group. If trust is an indication of the likelihood that one individual will engage in a costly cooperative interaction with another, then trust can be seen as the lubricant of social interaction. When trust is high, individuals are more likely to engage in potentially fruitful cooperative interactions with others in the expectation that they will reciprocate; conversely, a low-trust environment can induce individuals to forego cooperative interactions if they expect others to defect. There is even evidence that people in modern societies respond to environmental cues to effectively gauge how trustworthy other individuals in their local area are, and modify their tendency to engage in costly cooperative or altruistic behaviour accordingly (O'Brien & Wilson, 2011).

Another example where trust could change the outcome of behavioural strategies or norms is whether there is trust that others will conform to the normative system itself. If a normative system is in place, but conformity is low, then that would likely prompt others to also relax their commitment to the normative system. An anecdotal case of this is the observed difference in adherence to road rules in different countries around the world today. It is an understated marvel that in many Western industrialised nations many drivers will stop at a red light in the quietest hours of the morning, even when there are no other drivers on the road, and the chance of being caught running the red is negligible. If humans were rational agents, they would likely just run the red. Conversely, in many other countries, a red light is considered to be little more than a suggested course of action rather than an imperative to stop. However, relocate a driver from one country to another, and it is likely they will begin to conform with the local norms (or lack thereof) in short order. This appears to occur not through experience of punishment, thus altering the expected payoff of running the red, but rather through observation of other drivers' behaviour. A similar moral example concerns the temptation to engage in corruption and graft. Even when individuals agree that corruption is harmful, if they perceive that corruption is rife around them, they appear to be more likely to engage in corrupt behaviour (Tavits, 2005).

One explanation of this behaviour is that an individual's conformity (or otherwise) with the local norms is modulated by their perception of norm conformity in others. If this is the case, it raises an interesting prospect that could serve as a hurdle for the introduction of new norms into a moral system. If the introduction of new norms requires that old norms must be eroded first, then that process of erosion could undermine norm conformity in general, thus making the implementation of *any* new norms more difficult. Perhaps this is one contributing factor to the chaos and debased behaviour that has often followed even popular revolutionary uprisings throughout history.

One complication that arises from internal environmental dynamics is that they are particularly fluid and sensitive to feedback. In fact, a norm that alters the behaviour of its adherents might impact the payoff of many other behaviours or norms. For example, following the game theoretic examples in chapter 7, consider a normative system that regulates behaviour in a cooperative interaction modelled by the IPD. A norm that implements a kind of *tit-for-tat* strategy – say an "eye for an eye" rule – might begin to drive up the frequency of cooperation in the population. Yet, if eye-for-an-eye manages to drive cooperation above a certain threshold, such that many members of the population come to expect cooperation rather than defection, it might create an environment that is amenable to the introduction ("invasion") of a newly innovated strategy such as *always cooperate* – say as "universal compassion" rule, as expressed in some religions like Buddhism. If universal compassion were innovated before eye-for-an-eye was introduced, it would not likely have been very successful. However, once eye-for-an-eye makes the environment amenable, universal compassion can invade. Yet, universal compassion would change the environment yet again, leaving it vulnerable to invasion by nasty behavioural strategies. Those individuals who always trust others and who are likely to engage in costly cooperative interactions, could be haplessly exploited by others with no qualms about defecting. In fact, before eye-for-an-eye changes the environment, many individuals might be disposed to defect, not out of any motivation to exploit hapless cooperators, but as a defence mechanism against the majority of the population, whom they expect to defect – a classic case of a Nash equilibrium state.

As game theory has abundantly demonstrated, a population of undiscriminating cooperators is ripe for invasion by nasty strategies, thus making it unstable. As such, the norms employed by a population will change payoff for various behaviours, and not always in predictable or desirable ways. And this just reinforces the notion that any one behavioural strategy or norm is not likely to function optimally in every environment, as suggested by Axelrod and Hamilton (1981).

As with the external environment, environmental heterogeneity can further complicate the adaptive process. If it is difficult to gauge the state of the environment, it becomes more difficult to select an optimal trait or behaviour in that environment, and more difficult for a norm promoting a fairly narrow behavioural strategy to become established. Variation across the environment, or change over time, can thus increase the benefit of behaviours or norms that improve the quality of information about the state of the environment, and/or increase the benefit of behavioural flexibility or conditional strategies. This might be one of the factors that motivates norms concerning truth and honesty.

In this chapter I have underscored that the success or otherwise of any behavioural strategy – or norm that promotes it – will depend on the state of the environment in which it operates. This includes both the external and internal environment, both of which exhibit different dynamics. Much like organisms adapting to a complex world, moral systems also have to adapt to complex environments and weather the forces of heterogeneity and cooperative complexity. In the next chapter I will introduce a new and interesting complication: niche construction.

Chapter 11: Moral Niche Construction

> The bird a nest, the spider a web, man friendship.
> - William Blake

11.0: Building one's niche

In the previous two chapters I have elaborated the metaphorical notion of moral ecology, looking at how systems of moral norms adapt to their environment in response to the complex dynamics underpinning the problems of social living. Another factor that adds to the phenomenon of social and moral adaptive complexity is the active feedback from the individual or the group to shape its own adaptive environment. This feedback can change the very selection pressures that influence the evolution of the individuals or members of the group. While this process of "niche construction" is normally discussed in the context of biological evolution, I will apply it to the evolution of systems of moral norms. I will suggest that "moral niche construction" can effectively reshape the problem background to many of the problems of social living, changing both the nature of the problems and the optimal solutions to them, and enabling new solutions to enter a population and become dominant over time. I will argue that moral niche construction can also help explain one aspect of moral diversity, that of "ethical progress," or the progressive change of moral systems throughout history towards generally greater levels of compassion, tolerance and cooperation (Kitcher, 2011).

11.1: Niche construction

All organisms modify their environment merely by virtue of existing within it, if only by consuming resources, depositing waste and leaving their traces in the dust. And to the degree that this activity impacts fitness – say by overconsumption of a food source – this feedback between organism and environment is a relatively trivial observation. Typically, one might expect such activity to erode the capacity for that local environment to support the organism as resources are consumed, prompting it to either move on to a new local environment in the short term, or adapt until it sits in a rough equilibrium with that environment over multiple generations. However, organisms can also engage in a more direct interaction with their environment, actively shaping it in order to make it more amenable to their survival and reproduction, an observation originally made by R. C.

Lewontin (1983). The notion has since been developed extensively, particularly by John Odling-Smee, Kevin Laland and their colleagues under the moniker of "niche construction":

> Niche construction occurs when an organism modifies the functional relationship between itself and its environment by actively changing one or more of the factors in its environment, either by physically perturbing these factors at its current address, or by relocating to a different address, thereby exposing itself to different factors. (Odling-Smee, Laland, & Feldman, 1996)

One of the interesting effects of niche construction is that many organisms not only shape their environment to their selective advantage, but they then begin to adapt to their newly constructed niche. One paradigmatic example cited by Odling-Smee and colleagues is the spiders' web: the web may have originally been an adaptation to help catch prey, but once web-building is widespread it represents a new environmental niche, with new features and characteristics different from its former niche on the limbs of plants. And this new niche presents new adaptive challenges, such as camouflage, communication and protection (Preston-Mafham & Preston-Mafham, 1996).

Another example is a species that begins to build burrows to evade predators, but then adapts to its new subterranean environment with novel traits suited for that particular niche. Furthermore, these adaptations might subsequently make it significantly less fit in its original pre-burrow environment. Other examples of niche construction abound, including beavers' dams (Naiman, Johnston, & Kelley, 1988), termite mounds (Hansell, 1984), earthworms altering the soil in which they live (Lee, 1985) or certain pine trees that accumulate oils in leaf litter thus promoting forest fires, to which they have adapted (Whelan, 1995).

Niche construction can have a number of effects on the adaptive environment and evolution of an organism or a species over a variety of time scales. In the short term, it can enable an organism to shape its proximate environment in a way that impacts its own survival and reproduction. Yet the effects of that environmental manipulation might last only as long as the organism, such as a nesting site that requires constant maintenance by its owner. However, an organism can also shape the environment in such a way that it impacts the fitness of future generations, or even entirely other species or populations, a phenomenon called "ecological inheritance" by Odling-Smee and colleagues (Odling-Smee et al., 2003). In this context, the organism is effectively constructing a more durable niche that persists over generations, allowing a more complex feedback between organism and environment. It is this phenomenon of ecological inheritance that is of most interest in this thesis, particularly when it is applied beyond biological evolution.

Ecological inheritance effectively alters the conventional view of evolution as being a contest, of sorts, between an organism and an environment, whereby a (relatively) static environment poses particular problems and those organisms that best solve those problems go on to outcompete their rivals. With ecological inheritance, organisms can short-circuit this process by influencing the very problems inherent in their environment. In a sense, the organism does not need to start from scratch in each generation when it comes to carving out its preferred niche.

11.1.1: Fitness landscapes

One way to think about ecological inheritance is by means of a visual metaphor of a fitness landscape. In this metaphor the landscape represents various possible states of a population at any given time within a given environment, with the peaks and valleys representing the relative fitness of those various states in that environment. The peaks represent states of higher relative fitness, and valleys of lower relative fitness. When the population reaches the top of a peak, at which point moving in any direction will represent a reduction in fitness, it will reach equilibrium in that environment – at least while that environment remains static.

This does not mean there might not be higher peaks nearby, but they will be separated by valleys that must be crossed in order to reach them. Populations are also rarely – if ever – represented as a single point on that landscape, as variation within a population will cause it to appear more like a fuzzy blob resting in and around a peak. Broadly speaking, the two forces that can drive a population off one peak and through a valley into another is either mutation, which extends the bounds of the population blob, or environmental change, which alters the landscape itself. If a blob comes to straddle a deep enough valley, with two concentrations on either side, it might eventually undergo a speciation event. If the environment changes, and the population no longer rests on a peak but on a hillside, each generation will generally favour those individuals "uphill," thus eventually driving the population onto a new peak.

When visualising this metaphorical fitness landscape, short term niche construction is much like the population building its own temporary peak, one which disappears after that generation has passed. Ecological inheritance is like changing the landscape in a persistent or permanent way such that it affects future generations.

11.1.2: Cognitive niche construction

Niche construction theory need not only apply to alterations made to the physical environment. Andy Clark famously made the case that a cognitive functionalist interpretation of the mind ought to broaden our conception of cognition to be something that can happen beyond the boundaries of the skull, and employed niche construction in this cognitive context (Clark, 2008). He defines "cognitive niche construction" as "the process by which animals build physical structures that transform problem spaces in ways that aid (or sometimes impede) thinking and reasoning about some target domain or domains."

Kim Sterelny has further developed this idea, using niche construction to help us understand the explosive cognitive evolution of our hominin ancestors over the past four to five million years by looking at the changes our ancestors made to their social, informational and physical world (Sterelny, 2007, 2012). Sterelny argues that hominin cognitive evolution was driven by a positive feedback loop between cooperation and cognition, fuelled by social learning and guided instruction, along with increasingly sophisticated tool use and manipulation of the physical environment. The increasing complexity of these various constructed and inherited environments placed a selective pressure driving the evolution of more complex cognitive faculties.

One of the most significant selective pressures that drove the evolution of the ever more advanced cognitive faculties in our ancestors was the increasing complexity of the social environment – in the form of monitoring relationships, reputations and status hierarchies, along with the innovation of social learning, which enabled complex information to be transmitted across and within generations (Byrne & Whiten, 1989; Dunbar, 2003a). And to the extent that this social and cultural environment was passed from one generation to the next, then so too were the new selective pressures that they generated, representing a case of ecological inheritance.

This is in addition to the modifications we made to the physical environment, such as through the cultivation of crops or the domestication of animals. Another vehicle for niche construction was technology, with increasingly sophisticated tools – and increasingly demanding cognitive requirements placed on the creators and the users of those tools – which were also passed on through generations. The invention of the spear, for example, likely had a dramatic impact on the kinds and abundances of food available to early humans, and that invention likely had a significant effect on their selective environment.

As Sterelny puts it, “hominin evolution is hominin response to selective environments that earlier hominin have made” (Sterelny, 2007).

11.2: Scaffolding

One way to look at niche construction – and ecological inheritance in particular – is that each change to the relevant environment effectively erects a “scaffold” upon which future changes can then be built. These future changes may have been impossible to achieve without the earlier changes having been inherited. For example, the invention of thrown or projected weapons such as spears and bows – and the techniques to build them – likely opened up rich new sources of nutrients in the form of large game that would otherwise have been unavailable or too risky to procure. It also would have transformed inter-group conflict (Bingham, 2000). This, in turn, may have favoured new traits, such as hand-eye-coordination, or those facilitating cooperative hunting and coordinated defence behaviours (Sterelny, 2012). As these traits spread through the population, new dynamics could emerge, such as coalitions armed with throwing weapons out-competing coalitions armed only with handheld weapons, thus favouring traits that further promoted greater coordination of group behaviour, and so on.

The invention of one technology can easily scaffold the invention of another – particularly the invention of new tools – whether by facilitating new activities that would have formerly been untenable or impossible, or by creating new problems to be solved. This does not mean every change will be beneficial, but enough may have altered the environment such that future beneficial changes were made possible.

Extending the application of niche construction even further, changes in the informational environment can scaffold future changes, such as the evolution of abstract and recursive language changing the way individuals gauge others’ intentions and alter their behaviour accordingly, perhaps enabling new levels of complex social interaction and coordination of shared intentions (Sterelny, 2012). Even further, certain cultural innovations might have scaffolded additional changes, such as the innovation of symbolic marking and ornaments shared within a cultural group coming to facilitate better identification of in-group and out-group members, enabling the emergence of new norms detailing how to treat one’s own in-group (d’Errico, Henshilwood, Vanhaeren, & van Niekerk, 2005).

11.2.1: Moral scaffolding

I would suggest that the same niche construction and scaffolding processes can be applied to understanding how normative systems change over time. The emergence of new behavioural norms can effectively change the social environment such as to alter the payoff of various behavioural strategies. This, in turn, can facilitate the innovation and spread of new behaviours and behavioural norms that might have been highly costly or impossible to sustain in the earlier environments. Norms can thus alter the social environment, effectively erecting a scaffold that makes the emergence and spread of new norms possible. This is of particular interest in the moral context when the scaffolding can enable greater levels of cooperation, a process I refer to as "moral niche construction." I suggest that this could be a key mechanism that has enabled group sizes to increase dramatically beyond the levels seen in the Pleistocene, aiding in solving many of the problems that have emerged as groups have grown, and facilitating the massively ultra-social and cooperative societies we observe today.

One can view this phenomenon through the lens of evolutionary game theory by looking at how the presence of various strategies in a population in games like the Stag Hunt or the Iterated Prisoner's Dilemma change the environment, thus enabling new strategies to invade. In both games, high levels of cooperation or coordinated behaviour are often difficult to reach from less cooperative equilibria. From a starting point at a non-cooperative Nash equilibrium – let's call it the "state of nature"[12] in a nod to Hobbes – it is virtually impossible to immediately bump up cooperation to high levels with any significant stability over time. From an evolutionary game theoretic standpoint, those strategies that allow for high levels of cooperation cannot invade the population when it is resting at an evolutionarily stable non-cooperative equilibrium (Maynard Smith, 1986). In the case of the IPD, mutual defection can be thought of as analogous to the state of nature, which is not only a Nash equilibrium, but also the largest basin of attraction (given random interaction in an unstructured population). If a new strategy of unconditional cooperation is introduced into this environment, it will likely perform poorly and eventually disappear from the population.

In the case of the Stag Hunt, Brian Skyrms has shown that hare hunting equilibrium is the valley into which the population is likely to sink unless a population starts off with a large proportion of stag hunters (75 per cent or more, to be precise), or is structured in such a way as to make it very likely for the rare stag hunters to interact with other stag hunters (Skyrms, 2004). If an individual attempts to hunt stag in this environment, they will likely

[12] By using this term I do not mean to suggest that such a "war of all against all" has ever existed. There is abundant evidence that our primate ancestors engaged in cooperative behaviour well above that represented by the Nash equilibrium state in the Prisoner's Dilemma (de Waal, 2006), although humans appear capable of cooperating on a vastly greater scale than any of our primate ancestors.

perform relatively poorly, and will be out-competed by hare hunters. It will thus serve the stag hunter to imitate either the most common behaviour or the most apparently successful behaviour, which in both cases is hunting hare. In this kind of population, hare hunting has the largest basin of attraction, and one that is almost impossible to escape. Thus pervasive hare hunting can be considered the "state of nature," it is a deep basin of attraction, and is evolutionarily stable to boot, with it being difficult to climb out and reach the smaller but more productive equilibrium of pervasive stag hunting.

How, then, is widespread stag hunting or mutual cooperation to be reached? The key appears to be that even if the population is unlikely to jump directly from hare hunting to stag hunting, or from mutual defection directly to mutual cooperation in one leap – because the environment at those equilibria is particularly hostile to the strategies of stag hunting or cooperating – it can reach the higher cooperative equilibria over a series of scaffolded steps that change the environment to ultimately be more amenable to stag hunting or cooperating. While cooperation is always fragile in these games – there is often an element of downward pressure that threatens to drive the population towards the broad non-cooperative basins – there are other forces that can ratchet levels of cooperation up to higher levels.

In the IPD, for example, new strategies can invade even the Nash equilibrium of mutual defection and change the environment, thus enabling other new strategies to invade and spread throughout the population. For example, strategies that are somewhat "forgiving" of defection, and which only defect as retaliation to defection against them, can invade even "nasty" populations (Axelrod, 1987). This could represent a few more tolerant individuals entering a community, demonstrating their willingness to engage in and maintain cooperation, yet still remaining suspicious and vindictive against defectors. Another kind of moral scaffolding is changing the network structure of the environment such as to increase the chances of cooperators interacting with other cooperators (Nowak & May, 1992), thus raising the cooperation to slightly higher levels. This could represent members of a population banding together into a localised area and selectively interacting with other known cooperators. Once levels of cooperation hit certain threshold points, new even "nicer" strategies are then able to invade, such as more forgiving strategies, including strategies that could not have invaded the original population of pure defectors.

Eventually, the social environment can be changed by the existence of norms such that more consistent cooperators can invade and spread through the population – a feat that would have been impossible at the Nash equilibrium state. Likewise with the Stag Hunt.

Skyrms shows that even relatively minor changes to the strategies played by individuals in the population and/or to the structure of the environment can lead to higher levels of stag hunting. Sufficient changes can even bump the population out of the hare hunting basin and into the shallower (but more productive) stag hunting basin.

This evolutionary game theory picture relies only on replicator dynamics, driven by a kind of genetic analogue (Axelrod, 1987) or via imitation of either common variants or visibly successful individuals. However, add the mechanism of punishment and the environment can change even more dramatically. As Boyd and Richerson have shown, punishment can potentially entrench *any* behaviour by changing the payoff matrix of those behaviours (R. Boyd & Richerson, 1992). Punishment effectively alters the social environment in a persistent way, making certain behaviours more successful than they would otherwise have been, thus altering the selection pressures that favour certain behavioural strategies or norms. Punishment can also entrench sub-optimal or harmful behaviours – a norm might emerge that punishes stag hunting, for example – but to the extent that it has enabled entrenched behaviours that promote cooperation, it can aid in erecting the scaffolding upon which future innovations can be built. If the punishment happens to give stag hunting an edge, for example, it can help nudge the population out of the hare hunting basin.

The scaffolding mechanisms mentioned above are capable of altering an environment in any number of ways, not all of which will necessarily be towards more cooperative states. What, then, drove cooperation up even in the face of the downward pressure that often threatens to drive populations toward non-cooperative equilibrium states? As mentioned above, a plausible mechanism was competition between groups that provided an upward pressure favouring cooperation within groups (Bowles, 2009; Sterelny, 2010). If it was the case that individuals in more coordinated and cooperative groups were able to out-compete individuals in less coordinated and less cooperative groups, then there would have been a selection pressure favouring the most cooperative groups and their members. This may have introduced a sufficiently strong upward pressure promoting cooperation that overwhelmed the downward pressure towards the non-cooperative state of nature.

11.3: Moral progress

This process of moral scaffolding might also help account for the phenomenon of moral progress mentioned in chapter 2. It appears that a general trend over the past several centuries has been for highly punitive forms of punishment to fall out of favour in most cultures, including *lex talionis*, torture, inquisition, severe restitution, corporal and capital punishment and the harsh treatment of prisoners in war, to mention just a few (Kitcher, 2011; Pinker, 2011; Westermarck, 1906, 1932). The systems of punishment implemented in ancient societies, such as those of Rome or medieval Europe, were flat out barbaric by today's standards. Why, then, was punishment so severe in the past? And why has the severity decreased in recent times? As Philip Kitcher points out, this represents a kind of moral progress that deserves an explanation, particularly if one does not subscribe to the moral realist metaethical view.

Doubtless, many examples of harsh punishment were, at best, the products of the clumsy practice of cultural innovation and evolution and, at worst, the products of deviant or ruthlessly self-interested minds. Many would have proven to be sub-optimal (i.e. unnecessary, harmful or costly) solutions to the ends of solving the problems of social living. However, it might be the case that some of the norms that promoted or enabled these harsh practices were just products of different environments – physical, resource, informational or social – from the ones we see today in contemporary developed liberal democracies. And it might be the case that some of these norms helped scaffold the norms that we employ in developed liberal democracies today. In a sense, some forms of harsh punishment might have been important – if not necessary – steps to get to where we are now. It might be the case that in an environment closer to that of Hobbes' state of nature, there is a significant cost to employing any behaviour other than defection, given the high risk one will be defected against. And in that environment punishment might be a highly effective mechanism for driving cooperation to higher levels. This might help explain the proliferation of highly punitive systems of punishment that exact a high cost for serious moral transgressions throughout the world until relatively recently. Perhaps the benefit of heaving the population out of the state of nature was greater than the cost of even a rather sub-optimal or highly costly system of punishment.

Even if highly punitive normative systems were important in enabling cooperation in early societies, how then can the general reduction in the severity of punishment throughout history be explained? Moral ecology suggests at least two mechanisms that could account for the reduction in the severity of punishment over time. The first relates to the increasing efficacy of policing efforts throughout history. If one considers punishment in the context of a cost-benefit analysis of the agent, with punishment placing a cost on wrongdoing (or defection, in game theory parlance), then the punishment must be sufficient to make defection a more costly strategy than cooperation[13]. However, it is not only the severity of punishment that increases the cost of defection, but the severity multiplied by the chance of one being *caught* and punished. If the chance of being caught is low, then the severity of the punishment needs to increase in order to keep the average cost of defection above that of cooperation. If, however, the chance of being caught is high, then the severity of punishment can afford to be lower.

The detection and policing of defection might not have been a major problem for smaller scale societies, where the network environment is more likely to have been such that interactions would often occur between individuals known to each other. This also means a relatively simple informational environment, making it easier to keep track of likely cooperators and defectors. As Sterelny puts it, "in a village, everyone knows who the bastards are" (Sterelny, 2010). Even in slightly larger environments, indirect reciprocity can be effectively regulated by mechanisms such as gossip to help keep defection in check (Dunbar, 1998).

However, once populations increased, particularly in the post-agrarian world, it became more likely that individuals were interacting with strangers about whom they knew little or nothing. In societies such as these, the opportunities to defect – and get away with it – are likely significantly higher. Furthermore, prior to the innovation of processes or third party institutions that managed punishment, the meting out of punishment was more likely to fall on the shoulders of the individual defected against. Given the cost involved in doling out punishment – particularly after being potentially weakened by defection – it is more likely that even known transgressions often went unpunished. In such an environment, with relatively low rates of detection, the punishment needs to be proportionately more severe in order to act as an effective deterrent. However, as new innovations and institutions improve the rate of detection and policing, it allows the severity of punishment to be reduced.

This is not to say the innovation of such processes or institutions *guarantees* that punishment will be reduced. Only that reduced-severity punishment becomes a strategy capable of invading that population. This might account for why levels of punishment appear to have been lower in hunter-gatherer societies – where interactions were more likely to be repeated and the informational environment was more transparent making policing easier – then rising as group sizes increased in early city-states, peaking in early

[13] Of course, this calculus needn't be explicit to the agent; even if they simply imitate the most successful and/or common strategies in their population, they will be inclined to avoid defection if it is seen to be a less successful strategy than cooperation.

large-scale civilisations – where interactions were more likely to be anonymous or one-shot with a more translucent informational environment making policing harder – and then reducing again as new practices and institutions managing policing and punishment were innovated in more recent societies.

The second process that could account for a reduction in severity of punishment in recent times is that of the psychological internalisation of norms. I suggest that a great deal of defection in and around the "state of nature" is *defensive*, in that it represents individuals defecting in order to prevent being made a sucker. Yet, particularly once the fruits of cooperation have been visibly demonstrated, it is possible that individuals would be more inclined to cooperate, particularly if they feel safe doing so. Once norms that reduce the rate of defection are in place, and individuals begin to *expect* cooperation, and effectively trust their neighbours, then it might reduce the rate of defensive defection. Individuals might then begin to internalise the norms, and conform not out of an explicit sense of fear of punishment or defection, but because their expectations have changed, and so too have the behavioural options they consider in any particular cooperative interaction.

This might account for the phenomenon mentioned earlier whereby individuals in many developed countries are prone to stopping at red lights in the dead of night with no other cars in sight. In such situations, the chance of misfortune or detection if they run the red is negligible, yet a remarkable number sit and wait none the less. This could represent a case of the internalisation of the norms of the road, which is made possible by a generally high level of conformity to the road rules, and an expectation that all – including oneself – will conform. This doesn't mean there are not opportunists (or uncannily rational individuals) who might run the red, but the very apparent fact that not everyone runs the red when the opportunity presents itself suggest some interesting internalisation is going on. If this internalisation makes defection less likely, it reduces the need for punishment as a mechanism to deter defection. The internalisation effectively alters the social and behavioural environment, thus allowing new behaviours and norms to "invade," including those with more optimal (i.e. lower) levels of punishment.

All else being equal, it is desirable to keep punishment levels only as high as necessary yet as low as possible. Punishment might be a powerful mechanism to promote conformity to new norms and drive cooperation to higher levels, but punishment is not without cost. There are at least three. The first is the cost imposed on the punisher, which was discussed above (Fehr & Gächter, 2002). The second is the cost imposed on the punished. Every time *tit-for-tat* is forced into retaliatory defection, the aggregate payoff for the group is lowered from that of mutual cooperation. If there is significant inter-group competition, such reduction in group effectiveness could even threaten the survival of the group. While a certain level of cost is necessary in order for punishment to function, too high a cost and the system may become inefficient. Third is the cost of misapplication of the rule governing punishment, both in terms of guilty individuals avoiding punishment and, more importantly, innocent individuals being mistakenly punished. The cost of this "noise" can be considered as a drag on the aggregate payoff for the entire population. The more punitive this system of punishment, the higher the burden it places on punishers and the punished, and the less precise its application, the greater the cost. One can readily imagine the cost to a population of an overzealous scheme of capital punishment, for example. An optimal moral system will solve the problems of social living at minimal cost by raising the cost of defection to be sufficiently high to deter defection without imposing an undue drag on the population.

Thus, in historically early moral practices, particularly coarse and punitive measures like *lex talionis* and severe physical punishment, might have acted as scaffolding to promote conformity to, and internalisation of, normative systems, particularly as populations grew to sizes where reputations could no longer be effectively monitored on an individual level. Yet, once norm conformity had reached a certain level, the cost of these punitive practices may have proven to be sub-optimally high, thus making those systems less productive than systems that maintained similar or higher levels of cooperation without requiring such costly punitive measures. A society, and the individuals therein, with higher levels of norm conformity and lower levels of punishment might then out-compete a more punitive society and its members.

The notions of cultural niche construction and moral scaffolding can be useful tools in understanding some of the dynamics of how systems of moral norms respond to and change the environments in which they operate. These processes can also help account for at least some of what Kitcher calls "ethical progress," or apparently progressive change over time in the moral codes employed by a particular population or culture.

Chapter 12: Moral Ecology In Action

> Even in the angels there is the subordination of one hierarchy to another.
> - Saint Ignatius

12.0: Two problems

The notion of moral ecology outlined in the preceding chapters is necessarily highly abstract and rather speculative. In this section I will apply the moral ecology framework to some examples of social and moral phenomena in order to better illustrate its dynamics. I will do so by focusing on two of the more common problems of social living that are likely faced in some form or another by most of the populations of humans who have lived. I will then explore the impact that the various facets of the environment have on these problems and look at some possible solutions, as well as how these solutions can then lead to new problems to be solved.

12.1: Coalitions and groupishness

The challenge of forming and maintaining coalitions is one of the foundational problems of social living, as mentioned in previous chapters. A coalition is a splendid response to competition over some resource, since a coalition of individuals has an advantage over other solo individuals, or smaller or looser coalitions, as has been demonstrated in a range of species (Bingham, 2000; de Waal, 1982). A coalition can be an effective equaliser when faced with the threat of a strong or dominant individual exerting their interests. A coalition can also mount a more effective defence when facing hostile outsiders or neighbouring groups. Once formed, coalitions can foster cooperation internally, enabling individuals in the group to achieve feats that would otherwise be difficult or impossible, such as bringing down large game or enabling a more efficient division of labour. Coalitions bring many advantages, but they also raise a number of problems that must be solved in order for them to operate effectively, and the nature of these problems changes in response to the environment.

A central problem is one of managing the interests of the individuals within the coalition and aligning them with the shared interests of the group, thus preventing conflicts of interest from being destructive to members or the coalition as a whole. One might imagine that variation in the coalition members' interests could either exacerbate or alleviate this problem to some extent. A coalition of like-minded individuals, and similar beliefs and attitudes – perhaps instilled by a similar cultural normative framework – is likely to be significantly easier to establish and maintain than one composed of individuals with disparate views and varying beliefs and values. Where individuals who already share a great many values and beliefs might require relatively few behavioural and moral norms to reinforce group identity, individuals with a more disparate range of values and beliefs might require more norms that encourage costly displays of group loyalty in order to maintain the coalition. A coalition formed by members of different groups or organisations, with less apparent or perceived overlap in interests, might also be more likely to employ costly techniques to reinforce their coalition. Costly signalling is one mechanism that can help to demonstrate a commitment in such circumstances (Zahavi & Zahavi, 1997). One can view practices such as hostage taking in ancient or feudal times, or public expressions of alliance in more recent times, as examples of costly signalling devices intended to reinforce a coalitional venture between groups with potentially disparate interests.

Another challenge faced by coalitions is preventing free-riding, where one or more individuals contribute less towards the group's ends yet seek or expect a benefit disproportionate to their input. In this case, variation in the cost of free riding on the coalition might change the manner in which the coalition is maintained. For example, in times of scarce resources, free-riding might impose a higher cost on the other coalition members, thus making norms against free-riding more punitive in those groups. It seems certain that the punishment for slacking off during the production of a joint university assignment is likely to be less severe than that of slacking off during the joint defence of the group when under lethal attack from the outside. Similar variation is to be expected contingent on resource scarcity in terms of punishments for direct violations of other coalition members' interests, such as through cheating or stealing. The high cost of such activities to small groups living a subsistence lifestyle might account for their typically higher levels of punishment for such activities, including practices like maiming, a perhaps surprisingly popular practice up until the 19th century, particularly in the early days of the North American colonies (Earle, 1896).

As mentioned above, one of the prime motivators for coalition formation during our species past was likely inter-group conflict (Bowles, 2006). As such, variation in the number, relative strength and hostility of neighbouring groups would have an impact on the importance of coalition formation and internal stability. If, for example, the coalition existed in an environment with such limited resources that neighbouring groups tend to be dispersed widely, then there would be less pressure on forming strong and close-knit defensive coalitions. If, on the other hand, the groups are under constant mortal threat from outside invaders, there would be a substantially stronger impetus to ensure close knit and coordinated defensive coalitions. This might be one of the factors that precipitated the impressive militarism of the many Greek city states around Herodotus' time. Variation in external threat (real and perceived) might, in turn, translate into variation in the behavioural norms that are implemented to regulate the coalition and reinforce group identity, or "groupishness." Coalitions facing higher levels of threat would be more inclined to reinforce not only in-group favouritism but out-group discrimination, including requirement for more costly displays of group identity. The phenomenon of military uniforms, for example, might be in part a product of a desire to institute a sense of homogeneity and conformity between coalition members in precisely those times when reinforcing group bonds is most important.

12.1.1: Group size

Another variable that considerably impacts how coalitions are maintained is the size of the coalition itself. Given the benefits of larger coalitions, particularly the competitive advantage they enjoy over smaller groups, one would expect coalitions to generally grow in size over time. However, there are not inconsiderable hurdles to overcome in increasing group size. As the group size increases, so do the proximate resource requirements. As does the complexity of the informational environment within the group, amplifying the challenge of keeping track of in-group members (Sterelny, 2007).

It is naturally easier to track the identity of other in-group members in a small group. However, somewhere above "Dunbar's number" of around 150 individuals, the informational environment becomes so complex that our evolved psychological capacities to track identities are insufficient to accurately monitor group membership, relative status and reputations (Dunbar, 1998). At around this point, other mechanisms need to be innovated to help individuals identify in-group members and their relative standing. This might start off by taking pre-existing cues, such as readily identifiable traits including clothing, ornamentation, manners, customs or dialects, and reinforcing them with norms, effectively turning the into more reliable signals of group membership. This could also be one of the primary mechanisms that promotes the emergence of costly signals of group identity, such as piercings or facial tattoos, or strong dialects (Richerson & Boyd, 2005; Sterelny, 2012). Cultural mechanisms and signals such as these would aid in the expansion of group size beyond Dunbar's number, effectively scaffolding further changes that enable groups to grow in size even more.

Yet, there are other problems that emerge as coalitions grow larger. Once groups do grow in size, there is the distinct possibility that the variation in interests, beliefs and attitudes will also grow. So too the likelihood of sub-coalitions emerging which might inspire greater allegiance than the larger coalition. Nepotism, which might be a useful mechanism for encouraging small scale coalitions and cooperation along extended family lines, can end up becoming a corrosive force in larger scale groups if it fragments allegiances and creates internal tension among sub-groups within a broader coalition. The existence of sub-coalitions might eventually threaten to split the larger coalition apart, a phenomenon readily observed at scales ranging from the workplace to the geopolitical arena. As such, new mechanisms could be innovated to further extent and entrench group membership, such as encouraging group members to actively identify with the group. Norms encouraging a daily pledge of allegiance, or punishment of non-conformists, or regular vilification of out-group members, can build upon the earlier norms and extend coalition membership even further.

It is perhaps no surprise that one of Jonathan Haidt's purported five moral foundations – that of ingroup/loyalty – covers a broad range of norms that appear to serve the primary function of facilitating coalition formation and maintenance. Haidt and Graham suggest not only that this is one of the five core foundations upon which morality is based, but that we also have an evolved psychology that is innately sensitive to concerns in this domain:

> The long history of living in kin-based groups of a few dozen individuals (for humans as well as other primate species) has led to special social-cognitive abilities backed up by strong social emotions related to recognizing, trusting, and cooperating with members of one's co-residing ingroup while being wary and distrustful of members of other groups. Because people value their ingroups, they also value those who sacrifice for the ingroup, and they despise those who betray or fail to come to the aid of the ingroup, particularly in times of conflict. Most cultures therefore have constructed virtues such as loyalty, patriotism, and heroism (usually a masculine virtue expressed in defense of the group). From this point of view, it is hard to see why diversity should be celebrated and increased, while rituals that strengthen group solidarity (such as a pledge of allegiance to the national flag) should be challenged in court. According to ingroup-based moralities, dissent is not patriotic (as some American bumper-stickers

> suggest); rather, criticizing one's ingroup while it is engaged in an armed conflict with another group is betrayal or even treason. (Haidt & Graham, 2007)

Coalition formation and encouraging "groupishness" is one of the core challenges of social living, and one for which there are manifold normative solutions. However, the environment – including physical, informational and social environment – impacts which behaviours and norms will be successful at facilitating coalition membership. Any particular norm might be highly successful in one environment yet be inefficient in another.

Most norms will likely impose a cost on the members of a group, not only in terms of individual behaviours they encourage – such as costly signals – but also in terms of the potential for the norm to "misfire" and produce a harmful behaviour. Then there is the added cost of punishment to enforce the norms, as mentioned above. A norm that encourages highly costly displays of group membership, such as facial tattoos, might be very effective in a context where the cost of misidentifying an individual is potentially lethal, but might impose an undue burden in environments where the cost of misidentification is low, or the benefit lent by in-group members pales in comparison to the cost imposed by out-group members. Other norms that encourage out-group vilification might also turn out to be overly costly when group boundaries expand such that the two groups might benefit from becoming allies or it might hamper potentially fruitful cooperation between them. Even in the case of coalition formation, moral ecology suggests a diversity of norms will emerge and be differentially successful in different environments.

12.2: Dominance and hierarchy

As mentioned above, the formation of a group of individuals with at least some common interests does not necessarily dissolve the individual interests of the members of the group. Given that one would expect individuals to pursue their interests, even when acting within groups, there will almost inevitably be some conflicts that emerge. This is further exacerbated by the fact that a group is typically able to secure more resources than might otherwise be garnered by individuals alone, raising another problem of how to distribute these resources among the group members; a solitary hare hunter need not worry about being expected to share his hare, but a stag hunter has to confront the prospect of dividing up the kill with her partner. The more resources there are to share, and the more specialists who are directly or indirectly involved in generating them, the more complex the problem of their fair distribution – or distribution that is *seen* to be fair enough to avoid conflict. Conflicts of interests within groups can threaten to flare up and cause harm to individuals or disrupt the stability of the group as a whole, or weaken it in the face of competition by neighbouring groups.

Another challenge presented by group living is the coordination of activities within the group. This might be relatively trivial for a small group, or one that engages in synchronous group activities, where all members work towards a common goal at the same time, such as vigilance against predators or collective defence. However, coordination is substantially more difficult for larger coalitions, particularly those with greater specialisation and requirement for careful coordination, and which operate in asynchronous activities where group members work towards a common goal at different times.

One solution to these problems is the establishment of dominance hierarchies, which can potentially benefit both the dominants and the subordinates in the group (Boehm, 1999; Pusey & Packer, 1997). While subordinates might relinquish some of their power to pursue their own interests – perhaps even including their reproductive interests – they may receive other benefits in return. One of the chief benefits of dominance hierarchies can include a reduction of intra-group conflict, which might benefit both dominants and subordinates. Such a phenomenon is readily observed in animal behaviour, with a classic study conducted by A. H. Guhl and colleagues finding less fighting and a greater number of eggs laid in groups of hens with stable hierarchies compared to those with unstable hierarchies (Guhl, Collias, & Allee, 1945). Frans de Waal also observed that when hierarchies in a colony of male chimpanzees were unstable the chimps would engage in five times more damaging fights (de Waal, 1982). One can imagine the cost to individuals and the group as a whole of such violence and instability, particularly if under pressure from external competition from other groups. To the degree that a dominance hierarchy yields a greater benefit than the cost it imposes on the aggregate interests of the members of the group, then it can represent an effective solution to coordinating group activities and regulating intra-group conflict.

The importance of hierarchies – and the norms that promote and entrench them – will likely vary with a number of environmental variables. In terms of external environmental variation, if there is greater competition from neighbouring groups, there might be a greater emphasis on efficient coordination of group activities, such as in times of conflict or potential invasion. This might be one of the reasons why organised militaries throughout history have tended to evolve towards greater levels of organisation and hierarchy, to the extent of implementing rigid rank structures, despite such structures being costly to maintain. Scarcity of resources and a high density of groups within a particular geographical area might amplify this effect by further encouraging inter-group conflict and placing a selective pressure on more organised raiding or defence parties.

In terms of the internal environmental variation, the benefit lent by dominance hierarchies would be expected to scale with the size of the population, along with the degree to which members of the group are capable of specialisation and the volume of resources the group is capable of producing. This might account for why early hunter-gatherer societies tended to be less hierarchical than later post-agrarian societies, as Richerson and Boyd explain:

> Until a few thousand years ago humans lived in relatively small, egalitarian societies with a modest division of labor. Typical tribes consisted of a few hundred to a few thousand individuals. Tribal leadership was informal and leaders had only personal charisma to secure commitments of others. An egalitarian ethos was typically well developed. Division of labor by age and sex was important, but family, band and village units were largely self-sufficient except in subsistence or political emergencies. After the domestication of plants and animals, beginning about 11,500 years ago, human densities rose substantially and the potential for an expanded division of labor grew. Beginning about 5,000 years ago, complex societies began to emerge. Hierarchical states arose to administer the increasingly minute division of labor. Families became dependent on the products of strangers for routine subsistence. Leaders came to have great and sometimes quite arbitrary authority to coerce common citizens. Stratification emerged, with elites having highly disproportionate access to power and wealth. (Richerson & Boyd, 2001)

The existence of dominance hierarchies also alters the internal environment in the sense of favouring new behavioural strategies that might otherwise have been less productive. Rather than confronting the dominant individuals head on in a battle for superior standing, it might turn out to be more prudent to avoid their attention. For example, in some fish species where the dominant males corner most of the mating opportunities, a “sneaky mating” strategy can invade the population, where some males develop phenotypes that resemble females, allowing them to evade the dominant males and access females for mating (Gross, 1982, 1996). This suggests a frequency-dependence payoff for some behavioural strategies: sneaky mating is made more successful in the presence of dominant males than it would be otherwise.

An alternative is to seek to ally with dominant individuals, even if the alliance is disproportionately to the advantage of the dominant individual. In our own species, we see schoolyard cliques that accrete around one or more stronger individuals, such as bullies. For a physically weaker individual, or one unwilling to accept the risks of confrontation with the bully, allying with them is likely a more prudent strategy than opposing them, even if it means acquiescing to their commands.

However, dominance hierarchies present their own challenges, not least of which is the risk of corruption by the leaders, given their expanded influence over the activities of the group. There is also a certain point where nepotism shifts from being a virtue – encouraging cooperation between family and personal allies – to being a vice – favouring less optimal individuals for key positions of privilege, power or influence. Smaller groups, with fewer resources (and, therefore, power) to accumulate and where the mechanisms of indirect reciprocity and reputation can keep self-interested behaviour in check, may not need explicit norms that work to counter corruption or defuse power struggles.

However, larger groups will likely require such norms, and require they be suitably policed, a phenomenon that is observed in many cultures (Richerson et al., 2002). This process of developing systems and institutions of policing in turn produces a new problem of how to guarantee the policing of the powerful is not itself corrupted, particularly if the powerful have influence over the manner of policing; even if it is in the interests of group members to relinquish some power to the Leviathan, how can they be certain that the Leviathan will not abuse that trust? The innovation of a separation of powers can be seen as a milestone in cultural evolution tackling just this pitfall of hierarchical living. Hierarchies also raise prospects of power struggles and competition between prospective leaders, potentially leading to damaging conflict and fights over who inherits the dominant positions.

The existence of hierarchies also introduces a new pressure for individuals to seek status, effectively creating a new virtual resource over which individuals can compete. While status itself is insubstantial, its influence can be profound, particularly if high status affords the benefits of a greater proportion of resources and/or greater reproductive opportunities. Conversely, the costs of low status can also be high (Earley & Dugatkin, 2010). The competition over status can, itself, generate new avenues for conflict within a group, increasing the need to create new norms to regulate how status is earned and spent.

Given the importance of dominance hierarchies in the efficient functioning of large social groups, it is perhaps little surprise that there has been a selection pressure on our species favouring psychological mechanisms that can both promote and regulate hierarchies (Charlton, 1997). To this end, there is illuminating evidence that *Homo sapiens* have some evolved psychological tendencies to care a great deal about status within dominance hierarchies, including a tendency to imitate other high status individuals (Richerson & Boyd, 2005). We also closely monitor our status, identify our and others' relative positions in dominance hierarchies, and tend to expend a great deal of our energies seeking status (Buss, 1999; MacDonald, 1998; Tooby & Cosmides, 2009). Some of these tendencies I will discuss in more detail in the next chapter, particularly in relation to individual differences in how these tendencies manifest within populations and how they influence moral and political attitudes. While to some significant degree *Homo sapiens* appears to be a status seeking survival machine, there is also evidence that we have evolved not only a propensity to generate, conform to and jockey for status within hierarchies, but also a propensity to counter some of the tendencies towards generating gross inequalities within groups (Charlton, 1997), perhaps reflecting the many trade-offs and dangers inherent in allowing dominance hierarchies to run riot.

It is perhaps also not surprising that one of Jonathan Haidt's five moral foundations relates specifically to issues concerning the maintenance and regulation of dominance hierarchies – that of authority/respect. Like the other foundations, Haidt and Graham suggest we have an innate sensitivity to issues within this domain:

> The long history of living in hierarchically-structured ingroups, where dominant males and females get certain perquisites but are also expected to provide certain protections or services, has shaped human (and chimpanzee, and to a lesser extent bonobo) brains to help them flexibly navigate in hierarchical communities. Dominance in other primate species relies heavily on physical force and fear, but in human communities the picture is more nuanced, relying largely on prestige and voluntary deference (Henrich and Gil-White, 2001). People often feel respect, awe, and admiration toward legitimate authorities, and many cultures have constructed virtues related to good leadership, which is often thought to involve magnanimity, fatherliness, and wisdom. Bad leaders are despotic, exploitative, or inept. Conversely, many societies value virtues related to subordination: respect, duty, and obedience. From this point of view, bumper stickers that urge people to "question authority" and protests that involve civil disobedience are not heroic, they are antisocial. (Haidt & Graham, 2007)

Social living offers many potential benefits. But it also raises its own profound challenges, such as how to minimise internal conflict and how to coordinate activity within the group. Dominance hierarchies represent one solution that has emerged to tackle these problems, albeit one rife with trade-offs and with a propensity to generate new problems. Cultural evolution is an effective process for establishing norms that can help solve these problems, but it is also prone to entrenching sub-optimal norms and to cultural inertia, whereby the normative system takes some time to adapt to changed environmental conditions. This might be one of the contributing factors to why some cultures have strongly articulated moral norms promoting hierarchy and inequality, such as through caste systems or encouraging reverence of elders or leaders, while others strongly promote equality and attempt to suppress the pursuit of dominance, and some have a mix of norms. As such, moral ecology would predict that norms concerning social structure, hierarchy and dominance relations would vary between cultures, largely influenced by environmental variation and constrained by the messy process of cultural evolution.

Chapter 13: Moral Politics

> I asked Paul if he could think of a single question, the answer to which would be the best indicator of liberal vs. conservative political attitudes. His response: "If your baby cries at night, do you pick him up?"
> - George Lakoff

13.0: Intra-cultural diversity

The previous chapters outlined the notion of moral ecology primarily from a cultural evolutionary point of view, arguing that the dynamic complexity inherent in solving the problems of social living in a diverse range of environments yields a wide range of solutions to those problems. I have argued that the often fallible process of cultural evolution can account for a great deal of the observed diversity in moral systems around the world and throughout history, and that not all of this diversity has been counterproductive to solving the problems of social living. However, while cultural evolution can help account for some of the variation in moral attitudes and norms *among* cultures, it is less able to account for the variation in moral attitudes and norms *within* one particular culture. After all, it is not uncommon for two or more individuals to express different judgements regarding a particular moral matter despite being enculturated under the same moral framework, as illustrated by the quote from Edward Westermarck in chapter 2.

In this chapter I will focus primarily on *intra-cultural* diversity in moral attitudes rather than *inter-cultural* diversity in norms. A moral attitude can be defined as being a disposition for an individual to approve or disapprove of a particular act or of a moral norm in general. This is not to suggest some kind of non-cognitivist interpretation of moral utterances, such as that the meaning of a moral utterance is exhausted by its expression of approval or disapproval of a certain act or norm. Rather, only an individual might be disposed to approve or disapprove of a certain act or norm however they choose to express that attitude. The reason for focusing on attitudes rather than norms is that a single population might ostensibly subscribe to a single set of moral norms but there might still be diversity within that population in terms of attitudes about individual acts, or even about the validity of some of the norms themselves. The abolitionist, John

Woolman, for example, belonged to a culture that endorsed slavery, although he came to develop – and eventually express – an attitude contrary to the prevailing norm. During the time of his dissent, he presumably hoped to change the attitudes of his peers in an attempt to ultimately change the norm itself.

In this and the following two chapters I will argue that the same moral ecological dynamics that contribute to moral diversity in cultural evolution may also have influenced biological evolution and contributed to variation in our species' psychological makeup. This, in turn, has contributed to variation in moral attitudes within populations. I will first draw on the discipline of political psychology to look at how variation in psychological traits can account for at least some variation in moral and political attitudes. In the following chapter I will give an evolutionary story of how the complex and heterogeneous environment in our past has contributed to the evolution of a wide range of psychological traits. I will then weave these two stories together to suggest they can help account for at least some of the moral diversity we observe particularly within cultures.

13.0.1: Two theses

I will be arguing for two broad theses along these lines, one weaker and one stronger. The weak evolutionary thesis maintains that the mind is an organ that evolved to produce adaptive behaviour; however, the selective environment was such that there was no single set of psychological traits that would reliably produce adaptive behaviour in every environment. As such, our species has evolved a diverse range of psychological traits and dispositions, including personality traits and cognitive styles, that vary from one individual to the next within a population. This diversity is maintained in the population via a number of evolutionary mechanisms, notably bet-hedging and frequency-dependent selection. This variation has contributed to our species' ability to live successfully in a wide range of environments. This psychological variation also influences an individual's moral attitudes and behaviours, although this impact is effectively a by-product of our evolutionary past.

The strong evolutionary thesis resembles the weaker thesis but places greater emphasis on the adaptive significance of the social environment on the evolution of our minds. It proposes that it was particularly the adaptive significance of social living – and the complexity of solving the problems of social living – that has been a significant contributor to the evolution of psychological variation within our species. As such, it maintains that variation in psychological traits within our species – and the diversity in moral and social attitudes this variation produces – is not just a by-product of evolution, but was itself adaptive. Thus moral diversity is not only to be expected in a species like ours, but it might even exist for very good evolutionary reasons.

While there is reasonably strong empirical evidence supporting the weak thesis, the evidence in favour of the strong thesis is less firm, making it more speculative. However, I am inclined to believe the stronger thesis is at the very least plausible, and one that may in time be supported by new evidence in evolutionary and moral psychology. This chapter will first explore the evidence for how psychological variation can influence variation in attitudes, and will then go on to look at the evolutionary story of how our minds came to be furnished with a variety of personality traits and cognitive styles, and will conclude by looking in more detail at the two evolutionary theses.

13.1: Worlds collide

What would the world be like if everyone had the same personality? How would people think and behave in such a world? Would such a world exhibit a similar degree of diversity in moral attitudes as we observe in our own? A tantalising insight comes from one study conducted by Bob Altemeyer in 1998 (Altemeyer, 2003). He hosted two sessions of the Global Change Game, which is a simulation of the world that explores a range of social, economical, environmental and political challenges that faced the world in the late 20th century. It is typically "played" by between 50 and 70 participants, who are assigned at random to imagine themselves representing 10 regions of the world, with economic, military and technological resources corresponding to their contemporary real-world counterparts. Over the space of a few hours, players must manage their regions' affairs and interact with other regions to tackle global issues, such as trade, climate change, refugees and armed conflicts. Nuclear war is also a possibility, with it wiping out the entire Earth's population if it occurs. Clearly, these are largely issues of social and political import, but many of them also feature a moral dimension.

Typically the game is conducted with a random assortment of players to ensure a diversity of personality types, although Altemeyer changed that element in this particular study. He recruited his first set of players in terms of their scoring in the upper quartile on a test for Right Wing Authoritarianism (RWA), a metric that measures a range of behavioural dispositions and attitudes towards authority:

> High "RWAs" are authoritarian followers who have submissive attitudes toward established authorities, show a general aggressiveness toward persons "targeted" by those authorities, and adhere tightly to social conventions... high RWAs have proven

> to be relatively submissive to government injustices, unsupportive of civil liberties and the Bill of Rights, supportive of the Experimenter in the Milgram situation, high shockers themselves in a “punish the learner” situation, punitive toward law-breakers, mean-spirited, ready to join government “posses” to run down almost everyone (including themselves), happy with traditional sex roles, strongly influenced by group norms, highly religious (especially in a fundamentalist way), and politically conservative (from the grass roots up to the pros, say studies of over 1,500 elected lawmakers). They also have remarkably compartmentalized minds, endorse a multitude of contradictory beliefs, apply a variety of double standards to their thinking on social matters, are blind to themselves, dogmatic, fearful of a dangerous world, and self-righteous to beat the band. (Altemeyer, 2004)

After being separated into groups to represent their various regions, and spending 15 minutes studying their particular economic and environmental circumstances, the next step is for the facilitators to announce: “whoever is going to be the leader of your region, stand up.” However, among the groups in the high RWA simulation, nobody stood up immediately, with five men and three women eventually nominating themselves as leaders, with the last woman being pushed to her feet by the other members of her group. These leaders are made the “Elites”, with control over the region’s finances and military, and are secretly told they can pocket money if they so choose, with the “winning” leader being the one at the end with the greatest wealth.

After a single abortive attempt at inter-regional collaboration, the Elites retreated to manage their own region’s affairs in isolation. An overhead projector, which could be used to make announcements to the rest of the world, was rarely employed. Even a global ozone crisis prompted no multilateral or bilateral conferences. Besides very occasional interaction between two Elites to discuss trade, most groups kept to themselves to deal with their own internal problems. This they did relatively harmoniously, if slowly. There were no wars, although the less wealthy and developed countries fared poorly. By the end of the simulation, 1.9 billion people had died of starvation and disease, which the facilitators considered an extremely high fatality rate for any simulation without armed conflicts.

Altemeyer then ran a second simulation the following night – with a slight twist. Among the high RWA participants, he seeded seven individuals who scored in the upper quartile of another array, this time measuring Social Dominance Orientation (SDO). Those who score highly on the SDO test:

> Seem to be relatively power hungry, domineering, mean, Machiavellian and amoral, and hold "conservative" economic and political outlooks. And they are mostly guys. I have suggested (Altemeyer, 1998) that they would also tend to become authoritarian ("dictatorial") leaders to whom high RWAs would submissively flock. (ibid.)

This meant that amongst the 55 high RWA participants for the second run, seven were *also* high SDO. As Altemeyer notes, the propensity for high RWA individuals to be followers and high SDO individuals to be leaders means the two scales have little overlap, with a correlation of only 0.2. However, even that tiny overlap means there are a small number of individuals who score in the top quartile on both scales. These "Double Highs" (DH), as Altemeyer describes them, "want to be dominators" and he speculates that they endorse the submission revealed by their high RWA leanings because they "like the idea of others submitting to them."

The timbre of the second run of the Global Change Game was set when it took only 12 seconds for all the groups to have self-appointed a leader. By contrast, in the previous nights it took 15 seconds for anyone to even raise their hand. Four of those self-nominated leaders were DHs, with a fifth DH shortly thereafter becoming the effective second-in-command in his region, following his leader around and participating in all high-level negotiations and decisions. Another DH later staged a revolution in his group, which was one of the few led by a non-DH, effectively becoming its de-facto leader part way through the simulation.

In contrast to the previous night, the Elites interacted frequently in the second simulation, engaging in negotiation and deal making throughout the night. The overhead projector got a workout, and the ozone crisis prompted a global conference, although no united effort to confront the crisis ever emerged. The increased negotiation and trade did ameliorate some of the world's resource challenges, although there was a noted lack of charity, help for refugees or loans to the impoverished.

Conflict and the threat of war was more pronounced in the second simulation, with some regions acquiring nuclear weapons and threatening less powerful nations, which subsequently failed to secure any support from other powerful nations. After the bout of bellicose military posturing, many Elites made an increased investment in arms, and there was the looming threat of nuclear war between two regions that was only prevented when the allotted time for the simulation expired. At the end of the night, 1.6 billion had died from disease or poverty, which is only 300 million less than the previous high RWA simulation, although had the simulation continued for another turn, there is a good chance nuclear war would have wiped out the entire world's population.

13.1.1: Questions

Altemeyer is the first to acknowledge that the scale of this study makes it of limited explanatory utility in terms of drawing broad generalisations about high RWAs or Double Highs. However, his experiment is illuminating in other respects relevant to the enquiry on moral diversity in this thesis. Firstly, the study is suggestive of the link between variation in personality and variation in social and political behaviour; far from being cool rational agents, a range of psychological factors appear to have a significant influence on our attitudes and behaviours in a social context. For example, one of the hallmarks of high RWA individuals is a preference for conformity and insularity, along with a high level of out-group mistrust or ambivalence. When a group was composed of individuals with this tendency, it is probably of little surprise that this tendency was reinforced and their attention was directed inwards rather than outwards, even if this made it harder for the group to solve many of the simulated global social problems they faced. As I will explore in more detail below, there are many other dimensions of personality and cognitive style that also appear to influence social behaviour besides RWA and SDO.

The second intriguing phenomenon to emerge from Altemeyer's study is how the diversity of psychological and cognitive traits within a population influences the social dynamics of the group. Altemeyer started his study by asking: "what would it be like if everyone had similar levels of some personality trait?" Likewise, one might ask, what would it be like if people had a variety of personality traits? Would this have an impact on how they behave in a social situation? Would it influence the moral and political attitudes they form about how others ought to behave in a social situation? The interaction between the high RWA individuals and the Double Highs hints at how a diversity of personality traits in a population can result in emergent social dynamics. In this case, the Double Highs acted on their dictatorial and Machiavellian tendencies to take charge of the group and exert their will on their in-group as well as the Elites from their out-groups. The high RWAs – even if they did not agree with the actions of their Double High leaders – fell into line behind them rather than voiced a contrary view. One might equally run many more simulations of the Global Change Game with different mixes of personality traits and see very different dynamics emerge, with radically different outcomes from the simulated population of the Earth. And one might be prompted to wonder how the diversity of psychological and cognitive traits exhibited by the general population informs social and political behaviour in societies today and throughout history.

What is of particular note in Altemeyer's experiment is that the personality and behavioural traits exhibited by high RWAs and Double Highs were not just trivial aspects of the individual's disposition, but these aspects of their psychology dramatically impacted their behaviour in a social context, what attitudes they held and influenced they way they tackled social problems. The differences between high RWAs and low RWAs in terms of conformity, preference for hierarchy and respect for authority et cetera might be expected to have a significant effect on the way they structure their social relations, which in turn would affect how they tackle the problems of social living. As such, it is plausible that variation in personality and cognitive style, and variation in the makeup of the population, could have an impact on the moral attitudes and norms that are adopted by individuals in that population. Perhaps such psychological variation might even account for some of the observed variation in moral attitudes.

Furthermore, there is intriguing evidence that many aspects of personality have high heritability, as I will discuss below. This means that a substantial degree of variation in psychology is due to variation in genes. The question this raises is whether this genetic variation is an accident of evolution, as might be the case with variation in eye or hair colour, or whether the psychological variation has an *adaptive* explanation. After all, some genetic variation within populations is not just the product of stochastic forces, but is the product of natural selection, such as the frequency-dependent variation in beak size among *Geospiza conirostris* discussed earlier. Perhaps some of the variation we observe in personality and other psychological traits – particularly those that influence social and moral behaviour – also has an adaptive explanation. I will present an argument in this chapter suggesting that this might indeed be the case, and will explore its implications on moral diversity.

13.2: Moral proxy

One might think that an investigation of how individual differences in psychological makeup influence moral attitudes would come from the moral psychology literature. However, while there has been tremendous progress in revealing the psychological mechanisms that influence moral attitudes, and their possible evolutionary influences[14], there has been relatively little attention given to the notion that individual differences

[14] See *Moral Psychology* volumes 1-3 (2008)edited by Walter Sinnott-Armstrong for comprehensive overview of recent studies.

contribute to variation in moral attitudes and behaviour. Most moral psychology studies tend to seek universal patterns that underlie moral judgements formed by normally functioning individuals within a particular environment. The research that does investigate the impact of individual difference on moral judgement has tended to focus on psychopathy or brain damage (Casebeer & Churchland, 2003; Huebner et al., 2009).

As I will discuss in the next chapter, the majority of evolutionary psychology studies also concern themselves with uncovering the universals of human nature, such as the cognitive modules (if any) that have been shaped by natural selection over generations and which reside in us all (Cosmides & Tooby, 2004; Sidanius & Kurzban, 2003). Relatively little evolutionary psychology literature addresses the influence of evolution on individual differences in psychology and cognitive function – although there are welcome exceptions (Bateson, 2004; Buss, 2009) – and even less on the influence that individual differences have on moral diversity and moral disagreement. Yet studies that probe individuals' responses to moral problems demonstrate that responses are seldom uniform.

A paragon example is a study by Joshua Greene and colleagues who investigated subjects' judgements in response to a classic moral dilemma:

> A runaway trolley is about to run over and kill five people. In the *switch* dilemma one can save them by hitting a switch that will divert the trolley onto a side-track, where it will kill only one person. In the *footbridge* dilemma one can save them by pushing someone off a footbridge and into the trolley's path, killing him, but stopping the trolley. (Greene et al., 2009)

They noted that previous studies (Cushman, Young, & Hauser, 2006; Greene et al., 2001) had found that "most people approve of the five-for-one tradeoff in the *switch* dilemma, but not in the *footbridge* dilemma" and sought to identify the cause of the variation in judgement between the two scenarios. Interestingly, they found there to be a significant difference in the perceived permissibility of an action based on the subjects' perception of direct action (i.e. a physical shove) or an indirect action (i.e. pulling a lever). They found that a direct action causing harm was more likely to be perceived as impermissible, even if the subjects acknowledged that the agent intended to ultimately save lives. On the other hand, an indirect action that caused similar harm was more likely to be perceived as permissible in otherwise identical circumstances.

While the study focused on the variation in subjects' responses *between* the two trolley dilemma vignettes, what was left unexamined was the variation among attitudes *within* each vignette. There was no single vignette that elicited a uniform judgement of permissibility, suggesting there was at least some variation in attitudes about permissibility, and some disagreement among the subjects about the degree to which any particular action was permissible. This disagreement in attitudes about the permissibility or otherwise of an action from within a group of similarly enculturated individuals deserves an explanation.

13.2.1: From ideology to psychology

While there is a dearth of research in moral psychology directly tackling the question of the significance of individual difference in personality and cognitive function on moral proclivities, there is another source of literature that could act as a proxy: political psychology. This field is primarily concerned with understanding why individuals identify with certain political ideologies rather than others, and how they come to form the political attitudes they do. The discipline emerged in response to one of the perceived shortcomings of political science, particularly as practiced in the late 20th century, culminating in the so-called "end of ideology" movement. The issue was that "ideology" in political science had tended to be defined in a fairly restrictive sense, typically as being couched as a coherent and interrelated bundle of beliefs, values and attitudes about the proper goals of society and how they should be achieved (for a range of related definitions see Adorno, Frenkel-Brunswik, Levinson, & Sanford, 1950; Erikson & Tedin, 2011; Kerlinger, 1984; Rokeach, 1968). As John Jost has pointed out:

> They conceptualize ideology as a belief system of the individual that is typically shared with an identifiable group and that organizes, motivates, and gives meaning to political behavior broadly construed. (Jost, 2006)

The difficulty comes in understanding how individuals come to subscribe to a particular political ideology, or whether many people could even be said to explicitly hold a political ideology at all. All it would require is a little vagueness or internal inconsistency of beliefs to disqualify an individual from possessing a genuine political ideology. However, even if such inconsistent beliefs run rife, many people are still motivated to engage in political behaviour, not least through voting.

This stance towards ideology stems from a tendency amongst political scientists in the late 20th century to take a "top-down" view of ideology, where a sophisticated political ideology was created by political elites and was disseminated to the general public, with its uptake bounded by the cognitive and motivational limitations of the recipients (Jost, Kay, & Thorisdottir, 2009; Zaller, 1992). Many social scientists also employed the rational choice theory of human behaviour, borrowing it from the world of economics, and attempting to make predictions about political behaviour under the assumption that people are motivated to pursue their interests or preferences and behave rationally, no matter how irrational they appear to be (Scott, 2000). This tradition was less interested (or even entirely uninterested) in looking at the messy complexities of human psychology, but rather abstracted these confounding variables away in an attempt to provide a mathematical calculus that could predict political behaviour. It enjoyed some successes (Hechter & Kanazawa, 1997), but many notable phenomena proved intractable to the approach, such as explaining why individuals bother to engage in costly voting (where voting is voluntary) when the chance of them casting the pivotal vote is negligible (Fowler, 2006).

Contrasting the top-down approach is the bottom-up approach, exemplified by political psychology, which seeks to explore the actual psychological motivations for political behaviour (Jost, Federico, & Napier, 2009). Top-down approaches do take note of how political ideology is received and absorbed by the general public, invoking notions such as the level of political engagement they have, their education and cognitive sophistication and their access to information, but bottom-up approaches introduce a greater number of psychological variables that influence the uptake of particular political ideologies and individual political attitudes, and provide a more detailed explanation of how these psychological proclivities interact with ideology to produce political behaviour, such as voting.

The bottom-up approach tends to use the term “ideology” in a different manner to those of the top-down school. Instead of an ideology being a sophisticated abstract theory about the organisation of society, the bottom-up approach uses the term in a looser sense, referring to the cluster of beliefs and values adopted implicitly or explicitly by an individual, regardless whether it is influenced by the political elites or cobbled together from individual experience. As Jost et al. state: “although it is clear that people are far from perfect in their use of abstract ideological concepts, most citizens can and do use a subset of core values or principles that, for all intents and purposes, may be considered ideological in the sense of being broad postures that explain and justify different states of social and political affairs” (ibid.).

The perspective offered by the bottom-up approach suggests that irrational behaviour is not uncommon when it comes to the formation of political attitudes. Instead of taking a rational choice model of human psychology, which strips away any messy irrational inclinations in an attempt to render broad generalisations about human behaviour, political psychology wades into the complex mire of the human mind. It postulates that many largely unconscious predispositions and cognitive styles – some of which are relatively immune to environmental influence – interact with situational and experiential factors to produce a melange of attitudes that coalesce to form an individual's political ideological beliefs. Many of these beliefs might be inconsistent or outright contradictory, or they may be fuelled by irrational attitudes and desires, but they ultimately predispose an individual towards particular top-down theoretical political ideologies, which in turn help to organise and structure the individual's political beliefs and influence their political behaviour. As I will argue below, this view of political psychology has some interesting overlaps with moral psychology, which raises the prospect that research in political psychology could help to shed light on some moral phenomena, such as moral diversity.

13.2.2: From politics to morality

The analogy with moral psychology is intriguing. As discussed in chapter 4, there has long been a tradition in philosophy to consider moral attitudes to be a product of internal reflection and rationalisation – a product of taking our usage of moral language and our internal moral deliberation seriously. However, such naive rationalism about moral judgements appears to be unwarranted, as persuasively argued by the likes of Jonathan Haidt (2001). Haidt points out that many individuals hold conflicting and often contradictory moral attitudes and can often have difficulty justifying their moral judgements. As with political ideology, despite there being many moral philosophies to which one might subscribe, it appears that these do not simply trickle down from the elite – in this case presumably either moral authorities or moral philosophers – to be absorbed in their entirety by blank slate rational agents.

Rather, Haidt argues that most individuals appear to form their moral attitudes via a bottom-up process, whereby their emotional responses and intuitions – some innate, many enculturated – interact with situational and experiential factors to produce a melange of attitudes that coalesce to form an individual's moral beliefs. While many of these beliefs might be inconsistent or outright contradictory, or they may be fuelled by irrational attitudes and desires, I suggest they ultimately predispose an individual towards particular top-down theoretical moral philosophies, which in turn help to organise and structure the individual's moral beliefs and influence their moral behaviour.

However, unlike moral psychology, political psychology crucially benefits from the assumption that there is already diversity in political beliefs and attitudes within a population, and seeks not to explain such diversity away, but rather explain why such diversity exists. One objection might be that there is also diversity in attitudes towards food preferences or style of dress, and findings about variation in these domains might have little bearing on morality. However, political psychology benefits from being concerned with a domain that already overlaps with the moral domain in many areas. Many of the attitudes being assessed in political psychology have a moral dimension, such as attitudes towards equality, fairness, punishment, the treatment of outsiders or individuals of a different socioeconomic class, attitudes towards war, and even highly morally charged issues such as attitudes towards abortion or euthanasia. To the degree that political attitudes overlap or are influenced by moral attitudes, then insights into the diversity of one might pertain to the diversity of the other.

Some moral psychologists have already drawn a clear link between particular moral attitudes and political attitudes, with Jonathan Haidt and Jesse Graham finding that self-identifying liberals and conservatives[15] rate certain moral issues – such as those concerning harm compared to those concerning authority – with a different level of importance, with liberals rating harm/care and fairness/reciprocity as more important than authority/respect, in-group/loyalty and purity/sanctity, and conservatives rating them all as similarly important (Haidt & Graham, 2009).

13.3: Politics and psychology

The field of political psychology is rife with studies that tease out the influence that personality and cognitive function have on the formation of political attitudes. This section will give just a few examples in the hope that some might hint at a similar connection between psychology and moral attitudes.

The first example was introduced at the opening of this chapter in Robert Altemeyer's study of individuals who scored highly on the Right Wing Authoritarian and the Social Dominance Orientation scales. This study comes from a long pedigree of political psychology studies dating back to *The Authoritarian Personality* (Adorno et al., 1950). This study, performed in the wake of the World War II, was concerned with better understanding of the spread of social discrimination, authoritarianism and, ultimately, fascism. However, instead of just looking at the theoretical nature of these phenomena, it sought to find why certain individuals were drawn to these beliefs, postulating that it had

[15] The terms "liberal" and "conservative" here are used in the context of United States politics in the late 20th and early 21st century, although elements can be abstracted to other political cultures, such as those in other Anglophone democracies.

something to do with psychology. While flawed in its methodology (Martin, 2001), *The Authoritarian Personality* had a lasting impact on political psychology, if only for providing the field with tinder to spark further discussion and debate, and for introducing many researchers to the notion that politics might be influenced by (bottom-up) unconscious motives and psychological predilections rather than being driven purely by (top-down) rational discourse or environmental circumstances.

A more recent example of work in the field explores the link between the psychological need to manage uncertainty and threat – real or imagined – and politically conservative attitudes (Jost et al., 2007). This study found that while most people are motivated to minimise uncertainty in their lives to some extent, there is considerable variation in the extent to which individuals experience uncertainty as aversive and how they choose to resolve the uncertainty. Those who exhibited the greatest need to manage uncertainty and threat were more likely to hold politically conservative attitudes. There is similar variation in tolerance of ambiguity and its opposite, a tendency to stick to dichotomous conceptions and hold attitudes dogmatically, with those who have a low tolerance of ambiguity and a leaning towards dogmatism being more likely to hold politically conservative attitudes (Frenkel-Brunswik, 1948). Another variable is the propensity to perceive the world as being a dangerous place, and to have a somewhat pessimistic assumption about human nature at large (Altemeyer, 1998; Duckitt, 2001). This was again correlated with the individual holding politically conservative attitudes.

Another aspect of psychology that has been studied is integrative complexity, which is termed a "cognitive style," and refers to how people tend to integrate and process information. An individual with low integrative complexity will tend to take a black-and-white view of issues and will employ simple evaluative categories to attitudes, such as "good" or "bad" rather than taking a more complex view considering strengths and weaknesses of one particular notion (Tetlock, 1983). High integrative complexity is correlated with liberal views, while low integrative complexity is associated with conservative views[16].

One source of variation that would, on the surface, appear to have little to do with politics is personality as measured by the Five Factor Model (the "Big Five"): Openness,

[16] An anecdotal example might be the difference between former United States President, George W. Bush, and the Democratic presidential candidate in 2004, Senator John Kerry. The latter was criticised by conservatives and conservative elements of the media for being a "flip-flopper" because of his well known propensity to change his mind on particular policies. However, liberals praised this tendency, noting that he changed his mind on the strength of new evidence or argument. They, in turn, criticised George W. Bush for being overly simplistic and black and white in his thinking, a trait praised for its clarity and decisiveness by conservatives.

Conscientiousness, Extraversion, Agreeableness and Neuroticism. However, Jeffrey Mondak has found that two of these variables appear to have considerable predictive power when it comes to political attitudes (Mondak, 2010). High levels of Openness – which is often defined in terms of a propensity to seek our novel stimuli, to engage in intellectual pursuits, to be creative and express a general intellectual curiosity – are positively correlated with liberal self-identification and liberal attitudes. High levels of Conscientiousness – which is often associated with being organised, dependable, punctual and self-controlled – are positively correlated with conservative self-identification and attitudes. The other three factors have either conflicting or inconclusive evidence in terms of correlation with political attitudes.

13.3.1: Worldview

The mechanism that links psychology and politics appears to be the way that psychology influences an individual's "worldview" – defined as their broad framework for perceiving and understanding the world, and imbuing it with meaning and value (Anson, Pyszczynski, Solomon, & Greenberg, 2009). Worldview can be considered as the metaphorical lens through which each of us views the world, playing the role of a filter as much as enabling us to focus on those things that matter to us. Differences in worldview, which can be unconscious and obscured from introspection, appear to heavily influence political attitudes (Duckitt & Sibley, 2009; Duckitt, 2001; Lakoff, 1996). Thus, if someone perceives the world to be a dangerous place, they are more likely to hold conservative views than someone who perceives the world to be a relatively safe place (Jost, Glaser, Kruglanski, & Sulloway, 2003). Likewise, if they perceive the world to be a meritocracy – i.e. a world where people reliably get what they deserve, both in terms of rewards and punishments – then they are more likely to tilt towards conservative attitudes (McCoy S & Major, 2007). This phenomenon also carries over to worlds constructed in the imagination via thought experiments, with individuals self-identifying as either liberals or conservatives exhibiting a noticeable tilt towards conservative attitudes when the world is presented as being either dangerous or a meritocracy (Mitchell & Tetlock, 2009). Yet, when the evidence as to whether the world is dangerous or a meritocracy is ambiguous, variation in attitudes re-emerges, and does so along the lines of the psychological variations mentioned above.

While an individual's worldview has a strong influence on their political attitudes, their personality and other cognitive proclivities have a strong influence on their worldview. For example, someone who has a strong fear response, or a naturally low tolerance for ambiguity, or is excruciatingly shy, is likely to have a very different experience when presented with the same situation compared to an individual who has a mild fear response or a high tolerance for ambiguity or is a screaming extrovert. Consider, for example, an individual with a strong fear response. This individual might experience the same environment – say, a poorly lit park at night – in a very different way to someone with a low fear response. Different aspects of this environment have greater salience to the more fearful individual, emphasising some features while others escape notice. This different experience might then, in turn, reinforce their perception that the world is not a safe place: that bush that rustles and causes them to jump; those youths who appear to be eying them in a threatening way; the unlit pathway that is filled with menace.

As such, the very same environment appears very different when viewed through the lens of a different worldview. Yet the experience of that environment can also feed back and influence the individual's worldview. This, in turn, might make an individual more likely to find certain political ideologies and attitudes more appealing if they conform to their worldview. For example, someone with a strong fear response might be more attracted to the more security-inclined approach offered by many brands of conservatism[17].

However, ideology also works to influence worldview, leading to a feedback that serves to reinforce the worldview associated with a particular ideology. Consider the penchant for conservative media to emphasise crime, thus contributing to an elevated perception (or misperception) among many members of the public that they are likely to become a victim of crime, and perhaps contributing to more conservative attitudes and voting behaviour (Davis & Dossetor, 2010). The strength of this effect is still poorly understood, and will likely remain so until specific empirical studies are conducted to put it to the test. However, the evidence from political psychology is very suggestive of the influence psychology has on experience, which in turn influences one's worldview, which goes on to influence their political and, I would argue, many of their moral attitudes.

13.4: Morality, politics and genes

We can see from the above that at least some of the variation in political – and, I suggest, moral – attitudes is due to variation in psychological traits. Yet, at least some of the variation in our psychological traits appears to be due to variation in genes (de Moor et al., 2010; McCrae & Costa Jr., 2003). Repeated twin studies have found that identical twins are more alike than non-identical twins in virtually every personality metric (Bouchard &

[17] An old joke suggests a conservative is a liberal who has been mugged. While anecdotal, it is suggestive of a deeper point about the influence of experience on worldview, and in turn on political attitudes.

McGue, 2003; Bouchard, 1994). Nicholas Martin, for example, has overseen dozens of studies finding identical twins are more alike than non-identical twins when it comes to a startling array of traits beyond personality, including anxiety (Webb et al., 2012), depression (Ripke et al., 2012), migraine (Mulder et al., 2003), drug use (Verweij et al., 2013), intelligence (Benyamin et al., 2013) and many others.

Thus there appears to be a link between genetics and psychology, and a link between psychology and politics. This raises the question: is there a link between genes and politics? At least one study has reported just such a link. John Alford and colleagues conducted a twin study that found monozygotic twins are more alike when it comes to a broad range of political attitudes, such as school prayer, property tax and censorship all the way to modern art and gay rights, than are dizygotic twins (Alford et al., 2005). This is further suggestive that at least some variation in political attitudes might be due to variation in genes. And where genes impact some phenotypic trait, there is often (although not always) an adaptive story to tell. Before exploring this link between genes and morality/politics, it is important to better understand the influence that evolution has played in shaping our minds. It is to this I will turn in the next chapter.

Chapter 14: Evolution of a Complex Mind

We are a way for the cosmos to know itself.
- Carl Sagan

14.0: Evolution of the mind

Why do we have minds? Perhaps to ponder the stars (Sagan, 1980)? Or to reflect on the emptiness of existence (Sartre, 1956)? Or maybe to contemplate whether zombies have minds (Chalmers, 1996)? There are a number of ways to answer this question, but one fruitful approach is to look at the brain – and the mind it embodies – through the lens of evolution. From this perspective, the mind serves the primary function[18] of enabling its possessor to respond to environmental complexity with appropriately adaptive behaviour (Godfrey-Smith, 1998). It is this activity – rather than pondering stars et cetera – that best explains the recent maintenance of the trait under natural selection.

Virtually all organisms are capable of sensing and responding to certain features of their environment towards adaptive ends. There are clear adaptive advantages for an organism that can tell predator from prey, or potential mate from inanimate object, for example. This task is not overly complex in relatively simple or homogeneous adaptive environments, such as those where the adaptively salient features are few and predictable. In these environments, evolutionary parsimony will likely favour relatively simple behavioural dispositions, perhaps backed up by simple behavioural rules or heuristics that require minimal tracking of environmental features. However, producing adaptive behaviour is significantly more difficult in complex or heterogeneous environments. As such, understanding the dynamics and complexity of our adaptive environment is important if one is to understand the nature of the mind that has evolved in that environment.

14.1: Environmental complexity

From a population point of view in idealised circumstances, one would expect a trait that lends its possessors a selective advantage to increase in frequency within that population until it reaches a point of fixation. The result is that populations tend to gravitate towards

[18] See chapter 6 for a discussion of function in an evolutionary context.

a relatively stable set of traits that lend the highest fitness, driven by the process of directional selection. Of course, in the real world, this does not mean that all members of a population will end up identical; the mechanics of genetics, epigenetics, pleiotropy, genetic drift, linkage disequilibrium, disruptive selection and many stochastic and developmental factors complicate matters in practice. Still, in idealised circumstances, one would expect all members of a population living in the same environment to closely resemble one another, and that is what we tend to observe in nature. This raises something of a puzzle when different members of a population are found to exhibit significant variation in traits, and have that variation persist over generations. If this variation is not the result of the machinations of the evolutionary process or of random variation, then it deserves an explanation.

One explanation can be found in the nature of the adaptive environment in which that species has evolved. In order for a trait to reach fixation, it needs to reliably produce greater fitness outcomes for its bearer. By "fitness" I specifically mean what is often called "classical fitness," which is a property of an organism and is the product of its survival and fecundity (Dawkins, 1982). Yet the fitness outcomes depend on how the trait performs in its particular adaptive environment. If that environment is relatively simple and/or homogeneous, an adaptive trait has a greater chance of reaching fixation, as did the smooth morph of *Pseudomonas fluorescens* in the shaken beaker discussed in chapter 9 (Rainey & Travisano, 1998). If, however, the environment exhibits significant levels of complexity or heterogeneity, then it might be the case that there is no single trait, or single behavioural strategy, that will perform well in every state of that environment. In that case, it is possible for multiple traits and alleles – which are gene variants at a particular locus on the genome – to be maintained in the population over multiple generations, as with the wrinkly spreaders and fuzzy spreaders in the undisturbed beakers (ibid.). It is environmental complexity that appears to underpin a substantial proportion of the diversity in psychological traits and dispositions that we observe in our own species, and which I suggest contribute to at least some moral diversity.

Before discussing how environmental complexity influences psychological diversity in greater detail, it is worth taking a moment to clarify some terms. Section 8.2 already discussed a definition of "environment" – or more specifically, "selective environment" – in an evolutionary sense, with it representing those features of the world that are relevant to an organism's or trait's fitness. Complexity is a somewhat more nebulous concept, although one that can be defined with sufficient fidelity for our purposes here. For simplicity's sake (if you will excuse the pun), I will adopt Peter Godfrey-Smith's definition of complexity as being best understood as *heterogeneity*:

> Complexity is changeability, variability; having a lot of different states or modes, or doing a lot of different things. Something is simple when it is all the same. In this sense, complexity is not the same thing as order, and is in fact opposed to order. Heterogeneity is disorder, in the sense of uncertainty. (Godfrey-Smith, 1998)

As such, "environmental complexity" is simply the conjunction of these two concepts, referring to the level of heterogeneity, disorder or uncertainty in an organism's selective environment.

There are two broad aspects of our adaptive environment that exhibit this kind of complexity, as discussed in section 8.3. The first is the external environment, which includes those features *external* to the population, such as the physical landscape and the climate. There is evidence that the Pleistocene – a geological epoch that lasted from around 2.5 million years ago to 10,000 years ago – was a time of relatively high climatic instability and variability (Finlayson, 2009; R Potts, 1996; Richard Potts, 1998). It may well be that this climatic variability – such as increased seasonality – was a key spark that started us on the evolutionary path towards bigger brains and greater cognitive powers, as is suggested by Kristin Hawkes and her colleagues (Hawkes, O'Connel, Blurton Jones, Alvarez, & Charnov, 1998). It is almost certainly the case that a changing climate influenced the evolution of our hominin ancestors in significant ways by altering a wide range of adaptive pressures, such as access to resources, vulnerability to predators, encouraging migration and the favouring of behavioural flexibility (Sterelny, 2012; Wrangham, 2009). Greater mobility and less reliable access to resources may have also increased levels of inter-group competition. This competition, in turn, would likely have placed further selective pressure on in-group cooperation in order to better compete against out-groups (Bingham, 2000; Bowles, 2006).

Migration can effectively be thought of as another form of climate change: as a population travels to a new region with a new habitat – perhaps as the result of being forced out of its original territory – it faces different climatic conditions to which it must adapt. Environmental change can also come as a result of the activities of hominin populations themselves. This could be in the form of physically altering the physical adaptive environment, such as through the depletion of resources or the domestications of plants and animals. It can also be in the form of environmental engineering via new physical and cultural technologies, which unlock new resources or create new competitive pressures.

As discussed in the last chapter, cumulative changes over many generations represent a form of niche construction that can significantly alter the adaptive environment, and can do so over relatively short evolutionary time frames (Odling-Smee et al., 2003; Sterelny, 2012).

The second aspect of our adaptive environment is the internal environment, which includes features *internal* to the population itself, including social dynamics, the traits and behavioural dispositions exhibited by other members of the group, along with the culture and technologies employed by that group. As discussed in chapter 8, the internal environment is likely to have exhibited far greater levels of complexity than the external environment. For one, many features of the social environment are strategic in nature, in that the payoff of a particular behavioural strategy will depend on the other behavioural strategies with which it interacts. These kinds of strategic interactions, as discussed in chapter 7, make for some highly complex and unpredictable dynamics. Thus, to the degree that the internal environment had a significant impact on our fitness, then these dynamics had an important role to play in the evolution of our minds (Sterelny, 2003). Furthermore, as discussed in chapter 11, the process of niche construction that enabled us to shape our physical environment to suit our adaptive needs can also apply in the social realm as well. New technologies, artefacts and cultural innovations enabled our ancestors to radically alter their adaptive environment, and did so over relatively short evolutionary time frames (Richerson & Boyd, 2005; Sterelny, 2012).

So important was the internal environment that it may have proven even more adaptively significant than the external environment in terms of how our minds evolved. After all, as Sterelny has observed, the last common ancestor (LCA) of humans and chimpanzees, which is believed to have lived around six to seven million years ago, is likely to have closely resembled today's chimpanzees (Sterelny, 2012). Yet, while chimps still closely resemble our LCA both physically and behaviourally, we both look and behave in a manner profoundly different. Perhaps the most dramatic transformation we have undergone is not in our physical but our mental characteristics, possessing cognitive capabilities that far outstrip even our most erudite primate cousins. Given similarities in external environment shared between our hominid cousins and our recent hominin ancestors, the difference that drove this profound expansion of our cognitive faculties appears to have been in the internal – primarily the social – environment.

The notion that social environmental complexity has been the primary driver of the evolution of our cognitive faculties often goes by the moniker of the Social Intelligence

Hypothesis (Dunbar, 2003a; Sterelny, 2007). Nick Humphrey, for example, has pointed out that the uniquely powerful cognitive abilities of hominins appear to have evolved primarily in response to the adaptive pressures of the social rather than the physical environment (Humphrey, 1976). This, in turn, appears to have triggered something of a cognitive arms race. As intelligence increased, so too did behavioural complexity, and so too did the difficulty of predicting the behaviour of others (Sterelny, 2003). This theory is supported by Robin Dunbar's research, who has argued that group size is limited by the number of relationships that can be monitored, which in turn is limited by neocortex size (Dunbar, 1992).

This emphasis on the internal environment is largely due to the tremendous adaptive significance of social living and cooperation. An individual who is adept at navigating the social landscape and reaping the rewards of social learning and cooperative endeavour would arguably have had a significant adaptive advantage over an individual who was more adept at navigating the physical landscape but less capable at maintaining social and cooperative relations.

14.2: Responding to complexity

A basic or invariant behavioural rule might serve its bearer well in a simple environment, where the payoff of such a behaviour is predictably favourable, but such a rule will likely be less successful in a more complex environment. Such a rule might produce behaviour that is optimal in one state of the environment but sub-optimal in another. And if there is a sufficiently high selective penalty for producing the sub-optimal behaviour – such as an increased risk of predation or decreased chance of procreation – then the genes that contributed to that behaviour are likely to diminish in frequency within the population. As Kim Sterelny puts it, "behavioural flexibility is needed in complex environments, for in such environments invariant rules have mediocre rewards" (Sterelny, 2003). However, if an organism living in a complex environment is able to produce behaviour that is more suited to various individual states of the environment, it may end up having a significant selective advantage over individuals of a more rigid behavioural nature. It is in this way that environmental complexity can select for behavioural flexibility, with there being a number of mechanisms that can maintain such flexibility within a population.

14.3: Bet-hedging

Possibly the most rudimentary form of phenotypic flexibility is commonly called either "coin-flipping" or, more commonly, "bet-hedging." This is a strategy that is particularly prevalent in organisms that have been forced to adapt to relatively high temporal variation in their environment. As the name suggests, bet-hedging is likened to an organism taking something of a gamble, or "hedging its bets," on what the future state of the environment will be, and either adjusting its phenotype or producing a diversity of offspring that are adapted for each likely state. The end result is a genotype that has a reduced mean fitness across all possible states of the environment, but which also has lower temporal variation in fitness (Childs, Metcalf, & Rees, 2010; Philippi & Seger, 1989). In the game theoretic terms discussed in chapter 7, bet-hedging is akin to a homogeneous population of individuals who each employ a mixed rather than a pure strategy. In a sense, a group of bet-hedging organisms is simple at the population level and more complex at the individual level (Godfrey-Smith, 1998).

One example of bet hedging has been observed in *Pseudomonas fluorescens*, the bacterium mentioned in chapter 9. An experiment was conducted by Hubertus Beaumont and colleagues where the hapless bacteria were placed in media that were periodically shaken. Over several generations mutations emerged that benefited the bacteria in one environment, although these tended to be disadvantageous in the other environment. Eventually a strain emerged that produced two different variants as offspring, each of which was more suited to the shaken or unshaken environment, even though both variants had identical genomes. While the bacteria were continually switching between the two environments, this bet-hedging strain out-competed its more specialised brethren. As the authors note, such bet-hedging enabled the bacteria to rapidly adapt to changing environments, and likely represents one of the first evolutionary solutions to life in constantly changing environments (Beaumont, Gallie, & Kost, 2009).

Could bet-hedging account for some of the observed variation in *Homo sapiens*? It is possible, but only to a degree. Bet-hedging can be a strategy that emerges in environments that change between a relatively small number of predictable states, but do so at a relatively unpredictable rate. The dramatic difference between bacterial growth medium when shaken and unshaken is perhaps an extreme case, and one that is less likely to occur in nature. It might be more likely when considering other dimensions of variation, such as wet and dry, hot and cold, or resource-rich and resource-poor environments.

One potential example of bet-hedging in our species is variation in basal metabolic rate between individuals (Mootha & Hirschhorn, 2010). A high metabolic rate releases more energy over a shorter time span, affording greater levels of activity, but this comes at the cost of greater "waste" when at rest, and a greater requirement for fuel. A low metabolic rate might be more “efficient,” particularly in times of scarce resources, but comes at the cost of lower available energy. The higher thermogenesis associated with a higher basal metabolic rate might also lend an advantage to individuals in cold climates. These could be cases of bet-hedging, particularly if the optimal levels of activity and/or heat generation vary predictably across various environments with sufficient unpredictability over which environment one (or more precisely, one’s offspring) might reside in. If that environmental unpredictability persisted for sufficient generations, it might not have been possible for evolution to settle on one particular phenotype, and may have instead selected for a bet-hedging genotype. This is, of course, speculative, and it is entirely plausible that variation in basal metabolic rate is the result of stochastic forces, or is the product of many genes that are linked to so many other functions that metabolic rate alone could not be isolated as an axis of selection.

The examples mentioned above have all pertained to physical rather than psychological traits, although it is plausible there might be instances of bet-hedging here as well. To the degree that the environment is prone to vary over time, and to do so in unpredictable ways, then evolutionary bet-hedging can occur. For example, given the variable payoffs yielded by engaging in risky behaviour in different environments – say, resource-rich wet years compared to resource-poor dry years – variation in risk-taking or foraging behaviour might be cases of bet-hedging (Wolf & Weissing, 2010). However, both frequency-dependent polymorphisms and plasticity (discussed below) might be even more powerful explanatory forces in these cases.

14.4: Genetic polymorphism

Another response to environmental complexity – and one that is far more plausible as an explanation for individual differences in personality – is a genetic polymorphism. Where bet-hedging represents a single genotype that is capable of producing multiple phenotypes at relatively fixed ratios, a genetic polymorphism represents multiple genetic variants that are maintained over time within a population, contributing to multiple phenotypes (K. Smith, 2002). In game theoretic terms, a genetic polymorphism is somewhat akin to having a heterogeneous population of individuals, each of which employs a pure strategy. Thus a polymorphic species is complex at the population level but relatively simple at the organismal level (Godfrey-Smith, 1998).

In previous chapters I have discussed the crucial notion of selection driving populations towards a point of evolutionary equilibrium with their environment, whether that be a population of organisms or of agents in an Iterated Prisoner's Dilemma. This equilibrium is a hypothetical point at which the traits of a population become relatively stable, and where the population is resistant to "invasion" by new traits. This process is relatively straightforward if the environment is simple enough to afford only a single equilibrium point in regards to a particular trait. However, many environments are not so accommodating, particularly social environments. As discussed in chapter 7, many social interactions are strategic in nature, with multiple equilibria that can support multiple evolutionarily stable strategies (ESSs). If there are multiple equilibria, it is possible for multiple traits – underpinned by multiple alleles – to be maintained within the population over time. The cactus finch, *Geospiza conirostris*, and its two beak morphs is a textbook example of a genetic polymorphism maintaining phenotypic variation within a population, and nature is riddled with countless other examples (P. R. Grant & Grant, 2002).

There are a number of mechanisms that can maintain a genetic polymorphism, although the ones of particular interest to us are those that occur in response to environmental complexity. One mechanism is disruptive selection, whereby selective forces favour extreme variants of a trait. Disruptive selection can only occur in diploid organisms where homozygotes have an advantage over heterozygotes. This is a process that can in some cases even lead to a speciation event. To give a familiar hypothetical example, say there is moth with a gene that has two alleles: A and a. Given the moth is diploid, each individual has two copies of the gene, which can produce three different traits: black (AA), grey (Aa) or white (aa) wing colours. If the environment contains trees with both white and black bark (a simple example of environmental complexity), against which the homozygotes (black or white wings) can hide from predators, whereas the heterozygotes (grey wings) are conspicuously visible on both, then selection will tend to favour AA and aa over Aa. Presuming the moths interbreed randomly, this disruptive selection can maintain both alleles A and a in the population (if there is more interbreeding within populations with either AA or aa, then it is possible for a speciation event to occur given sufficient time). In terms of our psychology, it is *plausible* that some extreme traits might be favoured over more moderate traits, although it seems more likely that the opposite is the case (a notion with which Aristotelians will likely concur). As such, it is theoretically possible that disruptive selection is operating on our psychology, although I think it unlikely.

Another mechanism that can maintain a genetic polymorphism is balancing selection, which is a term that actually represents two key processes. The first is the opposite of disruptive selection and is called overdominance, or heterozygote advantage. As the name suggests, this is where heterozygotes have a higher fitness than homozygotes. The most famous example is the case of sickle cell anaemia, where heterozygotes have a slightly "sickled" blood cell that does not produce harmful levels of anaemia while also being resistant to infection by the malaria parasite, whereas homozygotes tend either to have harmful anaemia or are susceptible to malaria (Feldman & Cavalli-Sforza, 1985). Overdominance is a fairly common phenomenon, although it is not typically a response to environmental complexity per se, but is rather a complication produced by the complexities inherent in genetics. If we are to explain human psychological variation as a response to environmental complexity, then we will likely have to look elsewhere.

The final mechanism – and the one I suggest is primarily responsible for maintaining psychological diversity in our species – is frequency-dependent selection, specifically negative frequency-dependent selection. This occurs when the fitness of a particular trait – or allele – *increases* as its frequency within the population *decreases*. This is the process responsible for maintaining the two beak morphs in *Geospiza conirostris*, as mentioned previously: as the frequency of one beak increases in the population, this places relative pressure on the food source best exploited by that beak; this, in turn, means there is more food available for the alternate beak size, thus increasing its fitness until a relatively stable equilibrium is reached.

Frequency-dependence is, in a sense, a direct response to environmental complexity, however the environment in this context can be either the external environment *or* the internal environment. This is because the fitness of many traits is not only dependent on the state of the physical environment, but also on the other traits that exist within the population, such as we saw with the Hawk-Dove game in chapter 7. This introduces a highly strategic element to balancing selection, such as some strategic interaction with an aspect of the external or internal environment. And, as discussed in chapter 7, strategic interactions can have highly complex dynamics, often resulting in multiple equilibria and supporting multiple ESSs.

One classic example of a persistent polymorphism can be found in a region of our genome that is directly involved in a strategic struggle: our immune system. The major histocompatibility complex (MHC) is renowned for being the most polymorphic region on our genome, with many MHC proteins having over 500 variants (DeFranco, Locksley, & Robertson, 2007), and this genetic diversity has been maintained over countless generations in us and other primates, rodents and other creatures. The fact that the MHC plays a functional role in protecting against pathogens gives a hint as to why this polymorphism might be so persistent. Pathogens represent a highly complex and dynamic facet of our selective environment. In fact, many pathogens evolve over time in response to our ability to prevent their intrusion. As such, there is effectively an "arms race" between pathogens and our immune system.

In this context it becomes clearer why having a population with a monomorphic MHC might produce individuals who are at a selective disadvantage to individuals in a polymorphic population. If the MHC remains static for long enough, there is an increased chance environmental pathogens will evolve to bypass its protections, particularly as pathogens typically evolve faster than we do by virtue of having far shorter times between generations. A polymorphic MHC represents something of a moving target, thus making each individual within that polymorphic population harder to infect by a single pathogen (Slade & McCallum, 1992).

More specifically, a rare MHC allele has a selective advantage, particularly before pathogens have had an opportunity to adapt to it. While the allele enjoys a selective advantage, we can expect it to increase in frequency in the population. However, this increased frequency makes it a bigger target: once a pathogen mutates to counteract the allele, the high frequency of that allele in the population facilitates the transmission of the pathogen between hosts. This, in turn, would be expected to drive the frequency of the allele back down, perhaps to the point where its frequency is insufficient to sustain the pathogen. New pathogens will then emerge targeting new alleles, and so the dynamic cycle continues.

Adaptive environments with this kind of a strategic element are known to maintain polymorphisms via frequency-dependent selection. And besides pathogens, the other aspect of our adaptive environment that is also highly strategic has been our social environment. As such, frequency-dependent selection is a powerful tool that can help explain some of the persistent variation in traits in our species, particularly psychological traits, as I will elaborate on below. However, before exploring polymorphisms in our psychological traits, it is important to discuss another mechanism that has evolved in response to environmental complexity.

14.5: Plasticity

Perhaps the most sophisticated response to environmental complexity is to ramp up the complexity of the organism itself, and give it the tools to respond to a range of environmental conditions on the fly. This is the power of plasticity. Consider the hapless bacterium *P. fluorescens* and its experimenter-induced heterogeneous environment mentioned above. One adaptive response is to take a punt on what the future states of the environment might be and hedge its bets by producing different specialist morphs for each one, hopefully at frequencies that correspond to the likelihood of that environment existing. Another response would be to have a variety of morphs that co-exist, each occupying its own niche and sitting in balance with each other via frequency-dependent selection.

However, consider a more sophisticated hypothetical example of *P. fluorescens*, one that possesses the ability to detect whether it lives in a shaken or unshaken environment. If it could then "choose" to produce offspring specialised for that particular environment, it might hold a selective advantage over its bet-hedging or polymorphic brethren. This would represent a case of developmental plasticity, where the organism is able to steer its development in response to environmental cues. An even more potent mechanism is general phenotypic plasticity, whereby an organism can respond to its environment on the fly, as it were, and often re-adjust as circumstances change.

Plasticity has obvious adaptive benefits, particularly in complex environments where, as Sterelny stated above, invariant rules have mediocre rewards. However, plasticity is not a trivial capacity to evolve, not least in terms of the equipment required to alter its morph along with the sensory mechanisms of sufficient fidelity to reliably detect the state of the environment. As such, plasticity tends to be adaptive only in environments particularly where four conditions are met:

> These are when (1) populations are exposed to variable environments, (2) environments produce reliable cues, (3) different phenotypes are favored in different environments, and (4) no single phenotype exhibits high fitness across all environments. (Ghalambor, Angeloni, & Carroll, 2010)

Yet, even with these conditions, plasticity is such a powerful response to variable environments that it is a common sight in biology, from aphids (Braendle, Davis, Brisson, & Stern, 2006) to sea moss (Harvell, 1998) to snails (Stearns, 1989). Yet the undisputed master of plasticity is likey to be *Homo sapiens*. Our triumph of plasticity comes not in the form of altering our physical morph as much as in altering our behaviour to adapt to complex environments (Godfrey-Smith, 1998). We have not only excelled at trial and error learning, but also have more sophisticated mechanisms of steering our behaviour towards adaptive ends, specifically social and "apprentice" learning (R. Boyd & Richerson, 1995; Sterelny, 2012). Add to this the process of niche construction, discussed in the last chapter, and humans have been able to not only shape our phenotypes to our environment, but also shape our environment to influence our phenotypes. This is truly plasticity writ large.

Could plasticity explain some of the diversity observed among individuals in terms of moral attitudes and behaviours? Undoubtedly. Plasticity clearly accounts for a great deal of variation in attitudes and behaviours among individuals living in different environments or different cultures. After all, that is what plasticity is all about. It is likely that plasticity could account for a majority of diversity in moral and political attitudes *among* cultures. However, this chapter is particularly concerned with variation *within* rather than between cultures. Could plasticity play a role in *intra-cultural* moral diversity as well? I think it can, particularly in terms of how plastic organisms "decide" how to alter their behaviour. The very fact that plasticity is a heuristic process that relies on often unreliable environmental cues is also likely responsible for at least some of the variation in individuals' attitudes; each individual might be quite reasonable in their opinions given the way that individual perceives the environment, it is just that each individual sees a slightly different environment. This relates to the concept of "worldview" and its role influencing attitudes, as discussed in the previous chapter.

This effect might underscore one of the intriguing findings of Mitchell & Tetlock mentioned in the last chapter (Mitchell & Tetlock, 2009). These two conducted an intriguing study whereby they presented American subjects from across the political spectrum with three hypothetical worlds that varied in terms of levels of meritocracy, i.e. the correlation between effort and outcome in that world: one world had high correlation (0.9), one medium (0.5) and one low (0.1). The subjects were then asked to rank various income distributions for each of those worlds in terms of their fairness. For example, whether an egalitarian, efficient utilitarian or Rawlsian maximin distribution (Rawls, 1972) was most fair in a highly meritocratic world. The idea was to plumb the subjects' intuitions about fairness given whether the world was "just" or "unjust." Interestingly, they found that most subjects were in broad agreement about which income distributions were fair in the high meritocratic and low meritocratic worlds, regardless of their prior political attitudes:

> Both liberals and conservatives were willing to accept considerable inequality of wealth in high-meritocracy societies but with the reservation that distributions allowing people to fall below the poverty line remained unpopular for both ideological groups even in high-meritocracy societies... However, a majority of liberals *and*

> conservatives favored a Rawlsian "maximin" approach (Rawls, 1971) to the distribution of wealth in low- and moderate-meritocracy societies. (Mitchell & Tetlock, 2009)

So far most subjects were largely in agreement. However, when the subjects were presented with the moderate-meritocratic society – where it is most ambiguous as to whether outcome reliably flows from effort – attitudes split along pre-existing political ideological lines, with liberals advocating greater equality and conservatives favouring greater efficiency.

One interpretation of this result is that most people, regardless of their political persuasion, agree that some degree of welfare ought to operate in an unjust low-meritocracy world, but they simply disagree about whether *this* world is just or unjust. Thus, at least to some degree, political disagreement about welfare and income equality is not so much a disagreement in principle about when welfare ought to operate, but a disagreement in perception about to what degree our world is just or unjust. In more biological terms, this might be one consequence of the inherent translucency of the informational environment, with different people arriving at different conclusions about whether the world is just or unjust based on personal experience as well as other social and cultural factors that inform their worldview.

This is at least a plausible interpretation of Mitchell and Tetlock's study. However, it remains that many people still appear to be stubbornly resistant to reforming their *perception* of the state of the environment even in light of new evidence contracting their current view. This might be due in part to the fact we are highly *plastic*, but perhaps not quite so *elastic*. However, this story does not account for how individuals form their worldview in the first place, particularly individuals who are raised in very similar environments, perhaps even within the very same household. To explain how people arrive at their assessment of the state of the environment, we must turn to another aspect of plasticity, namely the constraints that are placed upon it by evolution.

14.5.1: Constrained plasticity

In biology as elsewhere, there is no such thing as a free lunch. So too with plasticity, and this can account for the limits that are placed upon it – limits that might help account for some of the variation in psychology that might underpin at least some of the observed moral diversity within cultures. Despite the adaptive benefits of phenotypic flexibility, plasticity often incurs costs not faced by organisms with more fixed phenotypes. One is the cost of developing and maintaining the faculties required to accurately track the state of the environment and steer the phenotype to be adaptive in that environment. Another typical cost is that it takes time and energy on behalf of the individual organism to track the environment prior to the development of an adaptive phenotype or the innovation of a new behavioural strategy. Functional complexity also often requires structural complexity, and structural complexity is typically more expensive to produce and more prone to breaking down.

Another potential cost stems from the difficulties in detecting the true state of the environment such that the phenotype can be adjusted to suit. Many environments are informationally translucent to some degree, with cues either being obscure, absent or lacking sufficient fidelity to represent their true state. For example, how is an organism to know it is sharing its local environment with a potential predator? There might be physical traces, remnants of previous prey, audible signals from other organisms targeted by the predator or other cues such as the presence of chemical traces, for example, odour. However, many of these cues will also be consistent with the predator having moved on to new hunting grounds. Adapting one's behaviour in anticipation of the threat of predation might involve sacrificing other fitness enhancing activities, such as spending time at a local watering hole. At some point, organisms have to take a punt and draw on the cues available to them to adjust their behaviour given the relative likelihood of these cues being accurate or not. This informational translucency is a particular challenge in social environments, as I will discuss in more detail below. These costs of plasticity mean that despite the adaptive benefits of phenotypic flexibility, there are often trade-offs compared to fixed phenotypes, or at least constraints placed on plasticity.

Plasticity is also fundamentally limited in practice: no organism can be infinitely plastic, nor is it likely that a hypothetical infinitely plastic organism would outcompete one with constrained plasticity. As Lars Penke and colleagues summarise,

> Even in the case of behaviour, unlimited plasticity is impossible to achieve adaptively, because the environment does not reliably signal the likely fitness payoffs for all possible behavioural strategies. In a complex world, environmental cues that can guide adaptive behaviour are inherently noisy, often contradictory, and unpredictably variable. The unreliability of environmental cues means that any behavioural plasticity based on trial-and-error learning must take time, because it must depend upon a decent sample of action-payoff pairings. Thus, given the complexities of real-world environments, organisms cannot instantly discern and implement the optimal

> behavioural strategy, so fitness-maximising by unlimited behavioural plasticity is an impossible ideal. (Penke, Denissen, & Miller, 2007)

As such, *ceteris paribus*, there will tend to be an adaptive pressure favouring specialists with lower cost simple traits rather than plastic generalists, particularly in relatively simple environments. Yet in more complex or heterogeneous environments, there will be an adaptive pressure in the opposite direction favouring plasticity. This tug-of-war can result in the evolution of limited or *constrained plasticity*.

However, the dimensions of the constraints on plasticity are not necessarily random. Indeed, the very constraints themselves can also be shaped by selective forces. For example, there might be certain behaviours that are likely to be adaptive in most environments, such as responding to putrefaction with aversion motivated by disgust. Yet even here plasticity can play a role, particularly in priming which stimuli trigger the aversive behaviour (Curtis, de Barra, & Aunger, 2011). We come pre-equipped with biases that are sensitive towards particular types of stimuli rather than others – including social stimuli – thus better enabling us to identify the proximate sources of putrefaction in our particular environment. There is still plasticity, but it operates over a constrained range of cues, emphasising some while deemphasising others.

Such constrained plasticity can be likened to having innate biases that favour some stimuli or responses over others in certain conditions, while still allowing for a great deal of behavioural flexibility. There is, in fact, ample evidence that we possess just such biases in our thinking, as pointed out by Martie Haselton and David Buss. Haselton and Buss have proposed what they call Error Management Theory (EMT), suggesting that many of these biases tilt behaviour towards adaptive ends, even if they also produce some irrational behaviour or beliefs in the process:

> EMT predicts that human inference mechanisms will be adaptively biased: (1) when decision making poses a significant signal detection problem (i.e., when there is uncertainty); (2) when the solution to the decision-making problem had recurrent effects on fitness over evolutionary history; and (3) when the aggregate costs or benefits of each of the two possible errors or correct inferences were asymmetrical in their fitness consequences over evolutionary history. (Haselton & Buss, 2002)

One example of EMT is the trade-off between type I errors (false-positives) and type II errors (false-negatives). If the cost of one type of error is greater than the other, then a cognitive system that is capable of making such errors might be biased in favour of one type over the other. For example, when detecting predators, the cost of a false-negative (not reacting to a predator when there is one there) is potentially far graver than a false-positive (fleeing a non-existent predator). A purely plastic organism, which would have to learn from trial-and-error the relative frequencies of type I and type II errors, could thus be out-competed by an organism biased to favour false-negatives. Haselton and Buss suggest EMT can account for why fear of snakes is so easily acquired relative to fears of other creatures, and why it is so difficult to extinguish once acquired. This is a clear example of constrained plasticity, with the constraints – in this case the biases – being themselves shaped by evolution.

In support of the weak thesis raised in the last chapter is the observation that much of our moral thinking appears to be likewise influenced by similar biases that emphasise certain features of the environment more than others in forming moral attitudes and judgements (Baron, 1994; Cushman, Knobe, & Sinnott-Armstrong, 2008; Haidt, 2001; Sripada & Stich, 2004; Tetlock, 2002). One example are the biases inherent in cultural transmission, such as a bias favouring cultural variants that are common or those that are possessed by high status individuals, as discussed in chapter 9, particularly if the cultural variants in question concern moral behaviours or norms. Another example is the finding that much of our moral thinking is driven by emotions that yield immediate intuitions of right or wrong rather than by dispassionate rational reflection. These intuitions are rapid, they emerge without conscious deliberation and they are strongly influenced by many biases, such as confirmation bias as well as biases leaning towards conformity and biases that avoid cognitive dissonance (Haidt, 2001).

A predisposition towards biases is just one example of how plasticity can be adaptively constrained. Another might be in varying the degree of plasticity expressed in individuals throughout the population. If there is no single optimal level of plasticity, there might emerge an element of bet-hedging in the amount of behavioural plasticity that is instantiated in the population. Some individuals might be more behaviourally flexible or sensitive to more environmental cues than others.

14.5.2: Personality as constrained plasticity

A pronounced example of constrained plasticity is the phenomenon of personality. Personality can be defined in many ways, but I will here refer to it as the broad behavioural tendencies that individuals exhibit, while leaving considerable room for behavioural flexibility. Many aspects of personality, while highly polymorphic, also appear to be highly heritable, meaning a substantial proportion of variation in personality traits within a population is due to variation in genes (Bouchard & McGue, 2003; Bouchard, 1994; McCrae & Costa Jr., 2008; Plomin, Owen, & McGuffin, 1994; Turkheimer, 2000). Furthermore, when surveying the political psychology research cited in the last chapter, personality appears to play a significant role in predicting an individual's political – and, I conjecture, their moral – attitudes. Perhaps the constraints placed on behavioural plasticity by evolution can shed some light on this connection between psychology and morality/politics.

Perhaps an unobvious question to ask is: why do we have a personality at all? What is its function, if any? It might seem so trivial that people vary in their behavioural dispositions as to be overlooked as a point of evolutionary interest. After all, random variation is the raw material for natural selection. Many evolutionary psychologists (such as Miller, 2000; Sidanius & Kurzban, 2003; Tooby & Cosmides, 1990 among others) have indeed downplayed the adaptive significance of individual differences in psychology, rather focusing on the "psychic unity of humankind" (Cosmides & Tooby, 1997). They have claimed that most individual differences are likely to be caused by forces other than evolution, citing *Grays Anatomy* and the fact that one can turn to any page and be confronted by features that will appear in every population around the world:

> There are strong reasons to believe that selection usually tends to make complex adaptations universal or nearly universal, and so humans must share a complex, species-typical and species-specific architecture of adaptations, however much variation there might be in minor, superficial, non-functional traits. As long lived sexual reproducers, complex adaptations would be destroyed by the random processes of sexual recombination every generation if the genes that underlie our complex adaptations varied from individual to individual. Selection in combination with sexual recombination tends to enforce uniformity in adaptations, whether physiological or psychological, especially in long-lived species with an open population structure, such as humans (Tooby & Cosmides, 2005)

Yet, far from being merely the product of stochastic forces, variation in personality may itself have an adaptive origin, contradicting the notion of an evolved "psychic unity" (Bouchard, 1994; Buss, 2009; D. S. Wilson, 1994).

One way to look at personality is as a *constraint on plasticity*. But not just any arbitrary constraint. In fact, the constraints themselves appear to be shaped by natural selection and are imposed across certain fairly discrete dimensions that are related to particular evolutionary trade-offs and frequency-dependent interactions. As Lars Penke and colleagues state,

> An evolutionary genetic conceptualisation of personality traits would thus be: individual differences in genetic constraints on behavioural plasticity, which lead to behavioural tendencies that follow individual reaction norms, and produce different fitness consequences in different environments. In short, personality traits are individual reaction norms with environment-contingent fitness consequences. (Penke et al., 2007)

Thus at least some variation in personality is not merely due to random variation or other stochastic evolutionary processes, but is maintained by selection (Wilson, 1994; ibid.; Ghalambor, et al., 2010).

For example, consider Extraversion, one of the traits in the five-factor model (FFM), or "Big Five," personality theory (McCrae & Costa Jr., 2008). Extraversion broadly represents an individual's tendency towards gregariousness and their propensity to seek social stimulation, along with a likelihood to experience greater reward responses from social activity. An individual who scores highly on the Extraversion scale will be likely to be outgoing, sociable, will maintain numerous friendships, participate in social and team activities, will be more likely to enjoy leadership roles and being the centre of attention. Those on the opposite end of the scale – i.e. Introverts – express many opposite tendencies, such as being socially withdrawn, reserved, will prefer solo activities, maintain fewer friendships and will avoid being the centre of attention (ibid.). Many twin studies have found Extraversion to have a heritability of between 0.4 and 0.5 (Bouchard, 1994). While this suggests that a significant amount of variation in Extraversion within the sampled populations is due to variation in genes, clearly there is a tremendous amount of variation from other sources, such as from non-shared environment.

How might something like Extraversion be an evolutionary constraint on plasticity? Daniel Nettle has argued that extraversion corresponds to an evolutionary trade-off between securing more mating opportunities (including poaching them from others) and higher risk-taking behaviour (including the risk of retribution from jealous rivals). He and his colleagues indeed found evidence that high extraverts today tend to have a greater number of sexual partners yet also experience higher levels of addiction and suffer more accidents (Nettle, 2005).

It is quite plausible that there was no single optimal behavioural strategy in terms of gregariousness and risk-taking that was successful at advancing fitness in every environment in which we evolved. And it might be the case that many evolutionary environments lacked cues of sufficient fidelity for an individual to make an assessment of the likely payoffs of the various strategies before they must commit to some behavioural path. Clearly we can expect *some* cues, and this could account for the substantial variability in Extraversion as a result of non-genetic factors. However, in such informationally translucent environments, there might be no single optimum along the Extraversion continuum that might afford selection gravitating towards a single point, resulting in balancing selection maintaining the continuum.

Another perspective is to think about traits such as Extraversion in terms of frequency-dependent selection (Penke et al., 2007). One might reflect on what the world might be like if everyone were an extreme extravert. It doesn't stretch the imagination to picture a world of intense social interaction, with a higher proportion of our cognitive resources devoted to maintaining those relations, concurrent with more transient mating relationships and more inter-personal conflict as a result. In such a world, an individual who represents stability and fidelity – and can effectively signal those traits – might possess a selective advantage over the more flirty and flighty Extraverts. Depending on the various payoffs, this dynamic could be cashed out as a Hawk-Dove game (see discussion in chapter 7) where the equilibrium state of the system is a polymorphism of the two strategies.

Nettle has also offered a similar frequency-dependent analysis of Neuroticism, with high levels lending greater vigilance towards dangers at a cost to long-term health (Nettle, 2006). Linda Mealy argues that frequency-dependent selection could also be responsible for maintaining psychopathy within a population at a low frequency (Mealy, 1997). Psychopaths in particular, with their glib superficial charm and other traits that endear them to unwitting victims (Hare, 1999), seem well suited to exploiting social interactions for their own benefit, reproductive and otherwise. However, above a certain threshold in a population, psychopathy could prove maladaptive, as the chance of one psychopath meeting another increases. In an evolutionary game theoretic sense, this is akin to an increase in nasty strategies within a population, which perform well when rare but perform poorly when interacting with other nasty strategies.

Richerson and Boyd also mention a frequency-dependent dynamic between innovators and mimickers in terms of cultural evolution, as discussed in the previous chapter (Richerson & Boyd, 2005). This represents another dimension that might influence individual differences in terms of novelty or a tendency towards conformity, perhaps corresponding to other FFM traits such as Openness to experience and Conscientiousness. And there are also many other studies looking at the various fitness trade-offs and frequency-dependent interactions that can maintain variation in personality in many species, not just our own (Dall, Houston, & McNamara, 2004; E. A. Smith, 2011; D. S. Wilson, Clark, Coleman, & Dearstyne, 1994).

14.6: Adaptive variation

Clearly people vary. That much is uncontroversial. A question raised above was: *why* do they vary? Clearly, a great deal of variation is due to environmental influence – responses to local circumstances, life history, experience, etc – which in turn is made possible by our evolved capacity for behavioural plasticity. Yet at least some of the diversity we observe in behaviour and behavioural dispositions can be traced to biological variation. This suggests a possible adaptive explanation – one we can find by looking at the strictures of our adaptive environment. As outlined above, there are very good reasons for thinking that at least some of the biological variation we see in our species' psychological traits was a response to the highly heterogeneous and informationally translucent environment in which we evolved – particularly the social environment. In the next chapter I will draw on this evolutionary story to revisit the weak and strong theses introduced in chapter 13.

Chapter 15: Evolution and Moral Diversity

Nothing is pleasant that is not spiced with variety.
- Francis Bacon

15.0: Evolved diversity

The last two chapters examined the question of whether evolution might shed some light on intra-cultural moral diversity. As discussed so far, at least some variation in political – and, I suggest, moral –attitudes is due to variation in personality and other psychological traits. I have also shown that at least some variation in psychology appears to be due to adaptive forces, particularly those imposed by our highly heterogeneous adaptive environment. However, just because some *A*s are *B*s, and some *B*s are *C*s, it does not follow that some *A*s are *C*s: it might be the case that the variation that evolution has maintained in our psychological proclivities is orthogonal to variation in our moral/political attitudes. Thus, what remains is to draw the final long thread from evolution all the way to our moral/political attitudes via the middle step of biology.

A strong indication that moral/political attitudes are influenced by biological variation comes from the study John Alford and colleagues mentioned in the last chapter (Alford et al., 2005; Alford, Funk, & Hibbing, 2008). They conducted an extensive twin study on the heritability of a wide range of political attitudes in order to get a sense of genetic and environmental contributions. While they found a great deal of variation in attitudes due to shared and unshared environment, as one would expect, they also found that the views of monozygotic twins correlated more strongly than the views of dizygotic twins. This indicated a degree of heritability across a broad range of issues such as school prayer[19] (heritability of 0.41), capitalism (0.39), the military (0.29) and the draft (0.38), foreign aid (0.35), individual rights (0.3-0.35), racial segregation (0.27) and the death penalty (0.32), with the mean heritability across all attitudes being 0.32.

The upshot of this finding is that variation in genes can account for at least some of the variability in political attitudes within a population. Given how polygenetic are the traits that influence political or moral attitudes, it is vanishingly unlikely that any single gene or

[19] This being a study conducted in the United States, where the question of whether prayer should be allowed in schools is considered a (rather hot) political issue.

cluster of genes for conservatism or liberalism will be found. However, it does lead to the startling conclusion that we are genetically predisposed to lean towards certain political attitudes, and quite possibly genetically predisposed to lean towards particular moral attitudes.

15.1: Two theses revisited

Thus we return to the two theses raised in chapter 13. The weak thesis states that the mind evolved to produce adaptive behaviour, but that the selective environment was such that there was no single set of psychological traits that could reliably produce adaptive behaviour in every possible state of the environment. As such, our species evolved not only a tremendous degree of behavioural plasticity, but also a diverse range of psychological traits and dispositions that constrain that behavioural plasticity in ways that vary from one individual to the next. It is this evolved psychological diversity that contributes to the differences in the way individuals perceive, experience and make sense of the world – i.e. the differences in their "worldview" – which in turn contributes to differences in moral attitudes and judgements.

15.1.1: The weak thesis

The weak thesis maintains that the moral/political diversity produced by psychological diversity is effectively a by-product of our evolutionary history, much as our penchant for sweet and fatty foods is likewise a by-product of a reward system that evolved to capitalise on rare opportunities to consume such energy rich sources of nutrients. As long as there is at least some variation in psychological traits between individuals, and this variation is at least partly heritable, and the psychological variation contributes to some of the observed moral/political diversity within groups, then the weak thesis is likely to be true. And there seems to be ample evidence to support the weak thesis, as outlined above.

For example, variation in integrative complexity might have been maintained within a population via bet-hedging or balancing selection, perhaps in response to an evolutionary trade-off between rapid but error-prone decision making and more costly and time consuming, but accurate decision making. Variation in the personality trait of Openness to experience might correspond to a frequency-dependent equilibrium between "innovators" and "followers." Until further studies are done to isolate the adaptive significance of the psychological mechanisms that underpin these psychological traits, there will remain an element of speculation about these claims. However, they are at least plausible. And if true, they lend support to the weak thesis.

That said, the very plausibility of the weak thesis risks making it somewhat trivial. Suggesting that variation in sporting performance corresponds to biological variation in basal metabolic rate, or VO_2 max (the maximal rate of oxygen consumption), or height, and that variation in these traits is due to past evolutionary trade-offs or frequency-dependent selection is hardly going to shake up the international sporting industry. However the weak thesis is just a stepping stone to the strong thesis, which carries greater significance for ethics and politics.

15.1.2: The strong thesis

The strong thesis acknowledges that our past adaptive environment has shaped our psychological proclivities, and that our minds have evolved to respond to environmental heterogeneity and complexity. However, it places foremost emphasis on the adaptive significance of internal environmental complexity rather than external environmental complexity. The thesis suggests that moral diversity as a result of psychological variation is not just a *by-product* of evolutionary forces, but the moral diversity was *itself* adaptive. Thus moral diversity is not only to be expected in a species like ours, but it might even have been beneficial in our evolutionary past.

The case for the strong thesis is underpinned by the apparent importance of the social environment to the evolution of our minds, as discussed in chapter 14. If the social intelligence hypothesis is correct, it was precisely the kinds of strategic dynamics that are modelled by game theory, as discussed in chapter 7, that have played the lead role in shaping our psychology. It was thus the runaway complexity of strategic social interaction inspired by the evolution of multiple levels of intentionality (Dunbar, 2003b) and driven by the fruits of cooperation that placed a powerful upward pressure on the evolution of more complex minds.

It is no accident that the social mind is often referred to as the Machiavellian mind (Byrne & Whiten, 1989; Byrne, 1996), as individuals seek to pursue their own interests within a complex web of other self-interested individuals, while navigating the complexities of cooperative interactions. Yet despite the sinister undertones of Machiavellianism, cooperation is ultimately the force that yielded the greatest potential rewards for our hominin ancestors. Thus, while pursuing self-interest, it is through cooperative interactions that one might best do so. Yet entertaining cooperation involves complex assessments of the likely behaviours of others, and braving the dynamics of strategic interactions like the Prisoner's Dilemma or Stag Hunt. Thus the same dynamics discussed in relation to moral ecology apply (Dean, 2012). And our minds evolved to parse these dynamics in the interests of maintaining prosocial and cooperative behaviour (Haidt & Kesebir, 2010; Hauser, 2006).

Thus, to the degree that variation in psychological traits was maintained because of the adaptive significance of strategic cooperative interactions, and to the degree that solving the problems endemic to these strategic cooperative interactions is one of the chief functions of morality or politics, then it is plausible that the kind of diversity we see in social behavioural strategies was itself a product – rather than a by-product – of selection. The variation in psychological traits within a population was maintained in part *because* of the variation it introduced into moral and political behaviour. Not only might an individual have been more successful by possessing a rare trait, but a population might have been more stable, coordinated and/or cooperative overall if balancing selection helped resist a drive to become a homogenous population exploiting only a single social behavioural strategy. And a more stable, coordinated and/or cooperative population can potentially be of benefit to many, if not all, of its individual members.

Consider Lomborg's IPD simulation, discussed in section 7.6. Lomborg found that, given a few reasonable constraints, populations engaging in repeated Prisoner's Dilemma interactions settled into state where a polymorphism of strategies was able to co-exist in relative (although not perfect) stability – what he calls a "meta-stable" state. The most commonly occurring meta-stable state, and the one that most impressed Lomborg, was the coexistence of what he called the "nucleus" and the "shield" strategies. The former were primarily trusting, forgiving and highly cooperative strategies that were highly productive when interacting with each other, but which were vulnerable to exploitation by more nasty strategies. The shield strategies were more suspicious, more punishing, and while they were less productive than the nucleus, they were more resistant to nasty strategies outside in the "wilderness." Thus the nucleus was protected from invasion by nasty "barbarians" by the shield in a state that represented a relatively robust compromise between productivity and resilience against nasty invaders. In evolutionary terms, there was a frequency-dependent relationship between the various strategies that maintained them in the population over time at relatively stable levels.

It is not difficult (perhaps perilously easy) to draw a parallel between the nucleus and shield and certain moral/political attitudes. For example, the nucleus resembles what might be called generally liberal moral/political attitudes that favour trust, cooperation and forgiveness whereas the shield resembles more conservative moral/political attitudes of suspicion, xenophobia and readiness to punish transgressors. It might be that the same dynamics that underpinned Lomborg's simulation have influenced the maintenance of psychological traits that predisposed individuals to either nucleus or shield strategies, or placed them somewhere on a frequency-dependent continuum.

Clearly, the strong thesis is more speculative than the weak thesis. However, given the adaptive significance of the social environment and of solving the problems of social living, it is at least plausible that the dynamics of these problems influenced the evolution of our minds. If this is indeed the case, then moral and political diversity within populations might not have been such a "bad" thing during our evolutionary past. It might have been the case that such diversity lent an element of "robustness" – in the game theoretic sense – to populations who were undertaking the "experiments of living", preventing them from driving too far into extremes when it came to solving the problems of social living.

If this is the case, then at least some of the moral/political diversity we observe even today within groups may be the result of evolutionary forces. Indeed, at least some moral/political diversity within groups may have proven adaptive in the past. However, I will save a discussion of the implications of such a thesis until the next, concluding, chapter.

Chapter 16: Moral Ecology and Diversity

> Without deviation from the norm, progress is not possible.
> - Frank Zappa

16.1: Moral ecology

If the primary function of moral systems is to help solve the problems of social living, and to help increase levels of social and cooperative behaviour, this raises a key question: which set of norms best satisfies this function? In order to answer this question, one needs to explore the devilishly complex "problem background" that moral systems face. To this end, I turned to evolutionary game theory in chapter 7, finding that the problems of social living, particularly those concerned with cooperation and coordination, are dauntingly complex indeed.

This is particularly because the success of any particular behavioural strategy, and the norm that promotes it, at solving a problem of social living depends on the environment in which it operates, both internal and external. The dynamic that emerges from this complex environmental dependence is what I call "moral ecology," as elaborated in chapter 8. Just as ecosystems evolve under the influence of manifold interactions and dependencies between and within species, with individual organisms and species carving out their particular niche, so too do systems of moral norms evolve. The cultural evolutionary process has seen individuals and populations innovate new behavioural norms that have ratcheted up levels of social and cooperative behaviour within their groups, given the contingencies of their particular environment. However, the vagaries of the cultural evolutionary process means this is often a clumsy and imprecise process. New norms are often sub-optimal, or downright counterproductive to the ends of morality. Yet, over generations, those systems that have proved stable enough, and have promoted higher levels of prosocial and cooperative behaviour thus benefiting their adherents, have tended to spread in favour of systems that have proven less effective at serving the function of morality.

However, moral diversity does not only exist among cultures but also among individuals within the same culture. Even two individuals enculturated within the same framework of norms might hold different attitudes on what they consider to be right or wrong, or which norms are the right ones to promote. It appears that the same complex dynamics that underpin moral ecology have also influenced the evolution of our minds, particularly our social and moral psychology, and this can help account for some of the moral diversity that appears *within* cultures. As I showed via an excursion into political psychology, it appears that many of our political – and, I conjecture, our moral – attitudes are influenced by aspects of our personality and psychological makeup. There is even evidence suggesting a link between genetics and political attitudes (Alford et al., 2005).

In chapter 13 I proposed two theses to account for the existence of the psychological variation that appears to influence variation in moral attitudes. The first is the weak thesis, which proposes that, given the heterogeneous environment in which our species evolved, there was no single set of psychological traits that would reliably produce optimal behaviour in every environment. As such, evolution was unable to settle upon a single optimal set of traits, or single set of constraints on our cognitive plasticity. The result is diversity in psychological traits, and this diversity influences our moral and political attitudes.

The strong thesis places extra emphasis on the complexities of the problems of social living, suggesting that the very dynamics that make morality such a devilish problem have contributed to the psychological variation that contributes to moral and political diversity. As such, variation in moral attitudes might not simply be a by-product of evolving in complex environment, but may have actually been adaptive, primarily by enabling individuals and populations to be more responsive to dynamic and changing social and environmental conditions, much like the genetic variability in our immune system lends individuals and populations an additional measure of resistance against environmental pathogens.

The ecology perspective is a powerful tool when it comes to understanding the presence and morphology of traits of an organism or species within a particular environment. It is also useful at explaining the differences *among* related species that exist in different environments, as well as in explaining patterns of diversity *within* a species living in the same environment. Likewise, I suggest moral ecology is a useful metaphor for understanding the presence and diversity of social and moral norms within a particular environment, as well as explaining the differences *among* cultures that exist in different environments, along with variation *within* those cultures. In this section I will review how the moral ecology perspective can be used to account for at least some types of variation that appear to exist both among and within cultures by looking at the various forces that contribute to moral diversity. As my intention is to provide a theoretical framework that might help explain some of the dynamics of moral diversity, the examples are necessarily abstract. However, they do suggest new avenues for research into the historical and anthropological literature to determine the extent to which these dynamics have been realised by living populations and cultures.

16.1: External environment

One of the major causes for variation in systems of moral norms is variation in the external environment, and the impact that environment has on the problems of social living. For example, variation in the amount of resources that are available can influence how large a social group can get, how it extracts and distributes those resources, how many other groups it might interact with, and how competitive those interactions will be. These factors, in turn, would be expected to influence many of the problems of social living and the moral norms that seek to solve those problems.

For example, environments with very poor resources would be more likely to have sparsely distributed populations that rarely interact. On the other hand, environments with richer resources are likely to afford a higher concentration of individuals – and possibly of groups – in one geographical range. This changes the nature of one of the key problems of social living, namely how one ought to interact with strangers, and the relative importance of behaviours that promote collective defence, raiding or trade, for example. This might influence moral attitudes and norms concerning trust, honesty, reciprocation and even which individuals qualify as moral agents worthy of protection from harm. Resource availability would also impact the relative benefit of implementing a strong dominance hierarchy, whether to reduce within-group conflict or to aid in the coordination of activities to extract resources from the environment. This, in turn, would be expected to influence moral attitudes and norms concerning things like respect for authority, loyalty and fairness of resource distribution.

Variation in climatic conditions can also influence some of the problems of social living. Environments with dramatic seasonal changes, or environments where the climate is particularly harsh at certain times of year, such as in high latitudes, would likely emphasise the challenge of efficient resource gathering at opportune times of the year, and emphasise prudent consumption when times are tough. Populations might be more likely to hoard food for the long winter months, introducing the challenge of defending that hoard against raiders. This might influence moral attitudes and norms concerning obligations towards group defence or responsibility not to over-consume stored food. Harsh environments might also account for moral norms that permit the euthanising of the elderly, a practice that would be unnecessary (and likely perceived as abhorrent) in more benign climates (see Sinnott-Armstrong (2006) for a metaethical perspective on such norms).

Note also that as cultures and technology develops, this in turn influences the physical environment by freeing up previously unavailable resources. The shift from hunter-gatherer lifestyle to a more settled herding and agricultural lifestyle also changed some features of the external environment, such as concentrating resources in highly developed areas (such as through damming and irrigation). Moral norms that might have been important in a hunter-gatherer society, such as those governing commitments to cooperative hunting or concerning the fair distribution of the quarry, might diminish in relative importance as the society becomes more settled. New moral norms relating to loyalty to the group and obligations towards collective defence might emerge and take precedence.

Thus, to the degree that the problems of social living are influenced by the state of the external environment, then we would expect there to be variation in the optimal solutions to these problems, and to the moral norms that seek to solve them. As such, we would also expect the cultural evolutionary process to gravitate towards different solutions, and different moral norms, in different external environments. This is likely to account for at least some of the variation in moral attitudes and norms among cultures, particularly if those cultures evolved in very different external environments.

16.2: Internal environment

While the external environment tends to be relatively static, and typically changes fairly slowly in response to the activities of the group over long periods of time, the internal environment – made up of the behavioural dispositions of other members of the group – is a far more dynamic entity. It can also have a tremendous impact on the problems of social living, and thus on which behaviours are successful within that environment. As such, it would also be expected to have a significant impact on the kinds of moral attitudes and norms that are adopted within a group.

An example of variation in the internal environment is the propensity for individuals to engage in costly cooperation with members of their own in-group. Such a disposition would be influenced by the amount of trust that members of the group feel towards other members of their own group. If trust is high, individuals will be more likely to engage in potentially fruitful costly cooperative interactions than if trust is low (O'Brien & Wilson, 2011). This may manifest in terms of moral attitudes and norms concerning trust, lying and cheating. For example, in an environment with low levels of trust, moral norms might emerge that place a significant cost on lying and cheating in the form of hash punishment for these actions. If these norms reduce the incidences of lying and cheating, and thus elevate levels of trust, then the cost exacted by the harsh punishment may begin to cause a drag on the population, thus enabling the emergence of new norms with lower levels of punishment.

In different internal environments, different behaviours will likely be successful at promoting cooperation, and other strategies will be less successful; it will likely be imprudent to engage in costly cooperative interactions in a population of likely defectors, for example. This kind of internal variation might account for variation in moral norms that promote cooperative interaction. A moral normative system that endorses a form of the *tit-for-tat* strategy, such as the "golden rule" or "an eye for an eye", might effectively dissuade defectors but at the cost of occasional costly cycles of mutual defection inspired by revenge (which might, in these cases, be called "justice"). Such cycles have certainly been documented in many cultures, including modern day Papua New Guinea (Diamond, 2012). However, such cycles might themselves change the internal environment, eroding trust and reducing the frequency of cooperation. An alternative strategy of promoting costly cooperation, such as moral norms that encourage highly trusting strategies (i.e. "love thy neighbour"), might avoid the cycles of mutual defection, but at the cost of being vulnerable to invasion by defectors and free-riders, which in turn might also erode trust and lower cooperation. Norms can also scaffold the introduction (or "invasion") of new moral norms, in principle enabling cooperation to reach new heights.

Another example of internal environmental variation that can affect the success or otherwise of various moral norms is the extent to which transgressions are policed and punished. In environments with poor policing, moral norms that promote high levels of punishment are likely to be more successful at promoting moral conformity than norms that are soft on punishment. However, when the levels of moral conformity rise to a certain threshold level, or if the level of policing improves, the highly punitive punishment might be more costly to maintain than less punitive punishments.

There is an element of frequency-dependence that complicates the internal environment, whereby the popularity of a particular behavioural strategy in an environment might make certain other strategies more or less fruitful. For example, in an environment of trusting cooperators, a willing defector will likely do rather well, at least in the absence of effective mechanisms to police and punish their behaviour. However, their behaviour will also influence the environment, changing it such that it might eventually favour alternative behavioural strategies.

Frequency-dependence is a powerful mechanism that can maintain variation within a population over time, as shown in chapter 7 with Bjørn Lomborg's "nucleus and shield" study (Lomborg, 1996). Lomborg found that a polymorphic population of different strategies in a sophisticated Iterated Prisoner's Dilemma simulation was often able to maintain higher levels of aggregate cooperation for longer periods of time than a monomorphic population all playing the same strategy. On an abstract level, one can imagine a population with a system of moral norms encouraging compassion, forgiveness and trust of outsiders as being vulnerable to defection or free-riding from exploitative individuals. Likewise, one can imagine a population with a system of norms encouraging retribution, harsh punishment for defection or free-riding and suspicion of outsiders as potentially losing out on many cooperative interactions with out-group members. Yet a polymorphic population that employs a mix of trusting ("nucleus") and suspicious ("shield") norms might be better able to resist the bulk of exploitation and defection, and also enable greater levels of fruitful cooperative interaction with outsiders.

This could be cashed out in the real world by looking for populations where there is persistent variation in moral attitudes or norms, particularly regarding cooperative interaction, such as norms concerning the scale of punishment for defecting or free-riding, or norms concerning trust or loyalty. And one need only look as far as the political pages in a newspaper from a contemporary liberal democracy to see such variation in action. It could be more than coincidence that what we might call "nucleus" norms seem to resemble those of modern day political liberalism (favouring trust and inclusion of out-groups, forgiveness rather than retribution against wrongdoing, a "dovish" foreign policy, and being "softer" when it comes to crime and punishment) and "shield" norms resemble those of modern day conservatism (favouring suspicion of out-groups, stronger retribution against wrongdoing, a more "hawkish" foreign policy, and "tougher" when it comes to crime and punishment).

The moral underpinnings of the liberal approach would likely make it more successful in terms of facilitating cooperative interactions that can benefit its members in a world of willing cooperators. However, the liberal approach would be more exposed to exploitation in a world of willing defectors. Conversely, the moral underpinnings of the conservative approach would be more resistant to the corrosive effects of defection in a world with a greater proportion of willing defectors. However, it would likely generate lower levels of aggregate cooperation in a world of willing cooperators due to its suspicion and the cost of harsher than necessary punishment. If the informational environment is transparent enough that one can reliably determine whether the world has a greater number of willing cooperators or defectors, then it might be relatively easy to know which moral attitudes and moral/political norms would be most appropriate. However, the social environment is rarely so informationally transparent. As such, individuals will have to effectively gamble on the state of the world and adopt a strategy to suit the world as they perceive it to be, which appears to be what people actually do (Mitchell & Tetlock, 2009).

Yet the tension between these two moral/political approaches when they co-exist within a single population might help to prevent any one approach from reaching fixation, which would make it harder to dislodge. This tension thus enables that population to more effectively respond to changing social environmental circumstances. Thus, perhaps the constant tension between moral/political attitudes in contemporary liberal democracies is a mechanism that enables greater levels of what Lomborg calls "meta-stability" and greater levels of cooperation over long periods of time. Moral and political disagreement, in this sense, is not necessarily a sign that something has gone wrong, but a sign that the population is capable of responding to a complex and changing environment.

Thus, to the degree that the problems of social living are influenced by the state of the internal environment, then we would expect there to be variation in the optimal solutions to these problems, and to the moral norms that help solve these problems. As such, we would also expect the cultural evolutionary process to gravitate towards different solutions in different internal environments. This is complicated by the high sensitivity of the internal environment to feedback, and by dynamics such as frequency-dependent selection, cultural evolution and cultural niche construction. This variation in the internal environment is likely to account for at least some of the variation in moral attitudes and norms particularly within cultures. It might also account for some of the variation in moral psychology among individuals, as discussed in chapter 14.

16.3: Functional equivalence

Sometimes two tools can solve the same problem equally well – or equally poorly. If morality is seen as a tool to help solve the problems of social living, then there may be more than one approach to solving a particular problem. There are likely to be many cases where there are multiple functionally equivalent solutions, much in the way the eye of a mammal and the eye of an octopus have evolved to have different structures but both perform very similar functions. Likewise, some of the observed variation among moral systems might simply be a result of the process of cultural evolution arriving at one of many solutions to a particular problem, while a different moral system has arrived at a different solution.

These functionally equivalent solutions might be rather similar, differing only in functionally irrelevant detail, or they might be very different solutions that are equally successful at solving a particular problem. An example of the former might be the difference in foods that are considered taboo in many cultures. A food taboo might be one effective costly signal of group membership, although the actual food that is prohibited is somewhat arbitrary to the function of the taboo. As long as group members visibly forego a certain potentially fruitful source of nutrients, then they have effectively signalled their group membership. There is little point in prohibiting a food that is not available in a certain environment, nor a food that no-one would wish to eat. But as long as there is consensus on which food is prohibited in the normative system, and people adhere to that prohibition, then the food taboo can function as a costly signal. Yet individuals from cultures with different food taboos might still be in disagreement over which food ought to be permitted and which ought to be prohibited. However, the disagreement over the content of the food taboo masks the fact that the norms serve the same function.

A related example of a divergent yet functionally equivalent class of moral norms might be prescriptions concerning dress codes, ornamentation or tattooing, particularly as costly signals of group membership (McElreath et al., 2003).

Despite the differences in justification that moral systems employ to account for the content of their norms, functional equivalence reinforces that many moral norms ultimately aim at solving the same problems of social living. And functional equivalence may help account for the variation in the content, if not the function, of many moral norms, particularly between cultures.

16.4: Sub-optimality

Like biological evolution, the cultural evolutionary process is not without its swings and roundabouts. Evolution works more as a tinkerer than an inventor, after all. As with genetic mutation, cultural innovation is often costly, and it is often easier and safer to make small variations to existing phenotypes or cultural variants than build new ones from scratch. As Richerson and Boyd point out, the process of cultural evolution is a powerful tool for enabling the spread of useful information and behaviours through a population, but it can also promote and entrench maladaptive information and behaviours as well (Richerson & Boyd, 2005). As such, one would expect that many of the moral norms that are produced via cultural evolution will not necessarily be optimised at solving the problems of social living in their environment.

For example, moral norms permitting revenge might turn out to be effective at punishing defection in environments with relatively low levels of cooperation, but they can also trigger costly cycles of mutual defection, i.e. feuds. This is particularly so in "noisy" environments, where mistaken defection is more likely to occur. When it comes to facilitating cooperation, a more optimal moral normative system might obviate revenge in favour of more effective policing, institutional punishment and/or more forgiving strategies that can more efficiently promote cooperation with a lower risk of cycles of defection. However, there is no guarantee that moral norms promoting these strategies will emerge in the population that adheres to norms permitting revenge.

Cultural evolution can also produce artefacts, such as cultural inertia, whereby it takes time for a culture to adapt to changing environmental conditions. Even if a new and more effective cultural variant is introduced, it will take time for it to spread through the population via imitation and guided variation. The process of punishment can also make experimenting with new (i.e. "deviant") behaviours more costly, even if those behaviours prove to be more optimal in that environment. As such, it does take some courage for individuals to change their moral system in the face of such inertia. There are countless stories of moral reformers experiencing stern resistance not only from cultural and moral authority figures but also the general population. Returning to the case of John Woolman and the movement to abolish slavery, even once his arguments were clearly articulated, there were powerful voices that opposed them and it took many years for the movement to enjoy broad support.

Phenomena such as cultural pleiotropy, where a cultural variant might have both positive and negative effects, can also introduce a cost to change (Keller & Miller, 2006). Introducing a new moral norm that offers a more optimal solution to one problem of social living might displace an existing moral norm that is more optimal at solving another problem of social living. Introducing norms that counter some of the negative effects of dominance structures can come at the cost of reducing the efficiency lent by hierarchical leadership roles, for example. Or new moral norms that ease stultifying conformity might come at the cost of enabling new and more disruptive cultural innovations to more easily invade a population.

Some moral norms also likely emerge to deal with perceived problems of social living, such as conflict over access to females of a reproductive age for example, and might offer sub-optimal solutions to those problems. One example might be moral norms permitting the corralling of women into harems controlled by high status males. That might, on the surface, reduce the amount of conflict between males for mates, but it might create new problems, such as reducing access to females for a majority of the population, generating new tension and demanding new solutions. A similar normative system would also likely favour the interests of high status males, who are precisely the ones with the power to resist change in the normative system that might benefit a greater number of individuals, both male and female.

Similarly, religion can serve as a potent mechanism for encouraging norm conformity, particularly through the mechanism of the "unseen enforcer" (Kitcher, 2011; Norenzayan, 2010). This is caused by the phenomenon whereby individuals are often more likely to conform to prosocial norms that place constraints on their interests if they believe they are being observed (Haley & Fessler, 2005). This appears to be due to a tendency to be mindful of preserving one's reputation and avoid punishment. Yet, there are many circumstances where one's actions are not likely to be directly observed. The challenge of encouraging norm conformity in these circumstances can be alleviated by introducing the notion of a deity or force which acts as an unseen enforcer, always observing and judging behaviour, often with the power to punish moral transgressions either in this life or the next. However, the religious framework that enables the invention of an unseen enforcer might in turn generate new problems, such as retarding the process of cultural evolution itself by stifling new innovations from entering a population.

One type of sub-optimality also stems from the fact that many of the solutions to the problems of social living are likely to be trade-offs. For example, moral norms that punish deviance, thus encouraging high levels of conformity, will likely stifle innovation, making it more difficult for the group to adapt to changing environmental conditions. Norms that encourage high levels of trust will likely make the population more vulnerable to exploitation by defectors. Norms that maintain strong dominance hierarchies can entrench inequality and disadvantage. Norms that allow for strict punishment will likely end up being highly costly when norm conformity increases, and so on. It is likely that for many problems of social living there is no single optimal point on the trade-off continuum.

The process of moral scaffolding described in chapter 11 can also show how it is unlikely for moral systems to rapidly advance to optimal levels. It is often the case that a number of changes to the internal environment need to take place before a new moral norm that promotes high levels of cooperation to be introduced into a population. It is upon these scaffolds that new norms can be erected, thus elevating levels of trust, conformity and cooperation over time. As groups expand, new norms are also required to regulate interactions in the changed social environment, and as groups increasingly interact with other groups, new norms are required to soften out-group bias and encourage out-group trust and cooperation. This means it is likely that a certain amount of sub-optimality will be expected in most cultures, which can contribute to some of the observed variation in moral systems.

16.5: Moral corruption

Another source of moral diversity emerges from the ability of those with power or influence to shape or preserve a moral system that operates in their favour. Moral norms that entrench power or privilege in a minority, such as norms demanding loyalty to high ranking individuals, or norms that strongly disadvantage some to the benefit of others, can be considered "morally corrupt." This is particularly the case when these norms do not effectively help solve the problems of social living faced by the group, or if they create new problems. As discussed in chapter 5, such norms might be deemed to possess the normative form of moral norms – i.e. by having apparent binding "practical clout" – and they might attract the same kind of condemnation and punishment when breached, as do norms that serve the function of morality. However, corrupted moral norms are simply not instrumentally effective at satisfying the primary function of morality. This is similar to how a law properly legislated can still be unjust. So too a corrupt moral norm might be adhered to and considered moral, yet it can fail to serve the function of morality. As long as the corrupt moral norms do not place an undue burden on the population, thus making its dissolution or extinction more likely, it is possible for such norms to persist over time. However, highly corrupt moral norms that are not suitably protected by those in power would be expected to be the target for replacement by newly innovated norms that better satisfy the primary function of morality, presuming such norms are innovated and given a chance to spread through the population.

The emergence of such corrupted moral norms is certainly not unexpected from the messy process of cultural evolution. Cultural evolution is such that transmission of new cultural variants is not guaranteed to favour those variants that are optimised to solve particular social or adaptive problems (Richerson & Boyd, 2005). A tendency to favour variants expressed by high status individuals can help entrench moral norms that benefit those high status individuals even at the expense of others. Furthermore, other moral norms can be introduced that stifle the innovation of new norms, particularly by employing punishment of behaviour that fails to conform with the existing moral system.

In fact, one way to frame Thrasymachus' error in Plato's *Republic* is as mistaking corrupted morality for genuine morality. Thrasymachus argues that "justice or right is simply what is in the interest of the stronger party" (Plato, 2003). While corrupted moral systems might have been prevalent throughout history, and while those individuals whom such moral systems serve might protect them vigorously, this in no way suggests they actually effectively solve the problems of social living. One feature of "moral progress" is removing such corrupted moral norms from within a broader system of norms, thus enabling greater levels of cooperation amongst a greater number of individuals.

Given the broad scope of possible ways a moral system can be corrupted, this can serve as another source of variation among cultures, at least in terms of what the members of that culture regard as moral. However, the dynamics of moral ecology primarily apply to those norms that do serve to solve the problems of social living, and are thus exposed to the complexities of solving those very problems. Moral corruption only serves to account for apparent variation, even if the norms concerned do not satisfy the function of morality.

16.6: Moral progress

One important form of moral diversity is the change that occurs over time in systems of moral norms. This is an especially interesting form of moral diversity particularly when this change over time appears to be progressive. Philip Kitcher offers several examples of moral change that appear to represent cases of "ethical progress", such as the move away from highly punitive eye-for-an-eye retribution norms, the shift from norms focusing on personal honour to those focusing on the common good, the change in the civil status of women, the repudiation of chattel slavery and the change in the perceived permissibility of homosexuality, among others (Kitcher, 2011). As Kitcher observes, without objective moral facts, such apparent moral progress risks looking like "mere change". However, here too moral ecology can shed some light on this form of moral diversity, particularly by looking at how moral systems change over time in response to their internal and external environments, and how each small change can potentially "scaffold" new changes that can better satisfy the function of morality.

With not inconsiderable difficulty, and perhaps driven by competition between groups (Bowles, 2008), humans have banded together in ever larger populations and innovated new solutions to the emergent problems of social living they face. In doing so, they effectively constructed new environments that scaffolded yet more innovations in ways of living and spawned new problems to be solved. As discussed in chapter 11, this process bears many similarities to the phenomenon of niche construction in biology, whereby individuals or populations alter their environment to such an extent that it alters the selective pressures that influence their further evolution. Likewise, a population can innovate new moral norms that effectively alter the social environment such as to facilitate the innovation of new moral norms that would not otherwise have been possible. For example, a moral norm that encourages harsh punishment for lying or cheating might enable the innovation of a new moral norm that encourages greater levels of trust between in-group members. It is this process of cultural niche selection – or "moral scaffolding" – that can ratchet up levels of cooperation over time. One of the consequences of this process is to enable larger and larger groups to form, which in turn changes the environment again.

Each level of human social organisation – from the band to the tribe to the chiefdom to the state – represents a different social environment, and each poses its own unique problems of social living not experienced by each level below it. For example, the problem of managing reputation is significantly easier to solve in smaller populations when individuals know everyone with whom they interact on a daily basis. The same problem is much harder to solve when the informational environment is more complex and translucent, such as when one lives in a mass society where many interactions are conducted with strangers. The problem of encouraging rule conformity is also significantly easier when there exists a trusted institution that polices those rules, and which incurs the cost of meting out punishment. It is much harder when rules are enforced, and punishment is meted out, individually on an ad hoc basis. The problems of preventing self-interest from generating conflict, or coordinating the activity of many individuals towards a common end, or fairly distributing resources, on the other hand, are fundamental problems at all levels of human social organisation with many potential manifestations that might vary among the various levels. Thus, as populations grow, they face new problems, and tend to innovate new moral norms to solve those problems.

For example, the harsher forms of punishment – such as *lex talionis* – that were common in Europe several centuries ago may have been effective at encouraging norm conformity and preventing defection, particularly if the rate of detection and enforcement was low. However, as norm conformity increased, and new practices and institutions were innovated that improved detection and enforcement, the cost of harsh codes of punishment began to overwhelm their benefit. Yet the increase in norm conformity that altered the environment can also scaffold the innovation of new, less severe punishment norms. If these new norms impose a lower cost and yet serve to effectively dissuade defection and socially disruptive behaviour, then they might become more widespread via the process of cultural evolution.

Thus, at least some of what Kitcher refers to as "ethical progress" is the process of moral ecology unfolding in different physical and social environments, with new behavioural strategies and norms scaffolding the innovation of new strategies and norms.

16.7: Psychological diversity

The above mechanisms account for at least a significant proportion of the observed diversity in moral norms among groups, although they are less effective at explaining diversity in moral attitudes *within* groups. As discussed in chapter 14, an individual's psychological makeup – their behavioural dispositions, personality, cognitive style, etc. – appears to influence many of their political attitudes. Given the overlap between the moral and the political domains, I strongly suspect that a similar phenomenon occurs in regards to moral attitudes. If this is the case, then psychological diversity could account for at least a proportion of intra-cultural moral diversity. I have suggested two theses that could explain why psychological variation could influence moral diversity: a weak and a strong thesis. Underpinning both theses is the notion that our minds evolved in order to produce behaviour in complex environments. While cognitive plasticity is one response to environmental complexity, plasticity itself has its limits. As such, plasticity is constrained, although it tends to be constrained along dimensions that represent particular evolutionary trade-offs.

The weak thesis suggests that the psychological variation that influences moral diversity is a by-product of these evolutionary forces. The more intriguing – and more speculative – strong thesis suggests that the environmental complexity that most influenced our psychological variation was the social environment. As such, the dynamics that underpin moral ecology may have influenced the very evolution of our minds, and might have contributed to psychological diversity not as a by-product, but because that diversity was itself adaptive. This psychological diversity appears to influence the way an individual experiences their environment, and thus how they form their worldview. For example, an individual with a strong fear response might perceive their environment as being more threatening and less safe, and they might then respond appropriately in terms of steering their behaviour in that environment. An individual's worldview then appears to make certain political ideologies, or moral norms or attitudes, more attractive to that individual. Once adopted, these ideologies or attitudes might then feedback into the worldview, reinforcing a certain view of the world. This might be one of the mechanisms that has contributed to the pronounced and often seemingly intractable disagreement between individuals from each end of the political spectrum in countries like the United States or Australia. Likewise, this mechanism might contribute to fuelling moral disagreements, such as that mentioned by Edward Westermarck in section 2.0.3.

16.8: Moral ecology and diversity

At the opening of this chapter I asked the question: which set of norms best satisfies the function of morality? As shown above, there is no single right answer to this question; there is no one set of moral norms that will reliably solve the problems of social living, and elevate levels of prosocial and cooperative behaviour, in every environment. In some cases there might be several sets of norms that satisfy the function of morality in a particular environment. In many cases the very existence moral norms will serve to change the environment in which they operate, and may alter the effectiveness of those norms or scaffold the introduction of new ones. Sometimes a variety of attitudes or norms working in concert – or even in tension – can best elevate cooperation to high levels. This is further complicated by the haphazard process of cultural evolution, and the possibility of innovating sub-optimal norms, along with the ever-present threat of moral corruption.

The upshot is that we should expect there to be diversity in the moral norms that are employed throughout the world rather than for there to be a single set of norms that all populations will gravitate towards. In the next chapter I will conclude by looking at some of the metaethical implications of this view.

Chapter 17: Conclusion

> If we cannot now end our differences, at least we can help make the world safe for diversity.
> - John F. Kennedy

17.0: Explaining diversity

In the absence of the light of evolution, biological diversity represents something of a mystery. Pick a geographical region and there are likely to be myriad creatures possessing innumerable forms and wildly divergent behaviours. Among locations creatures also vary greatly in number and composition. Even within a particular population there is a multitude of variations subtle and gross, such that no two members of that species are likely to be perfectly identical. Why does such profound diversity exist? And why does it take the form it does?

If such diversity were by design then the plan would appear to be well beyond human ken. Even if there was such a thing as a creature's prototypical form, it would appear from the vast array of variations that few, if any, embody that form in this world. It might even appear like a distracted creator was haphazard with its powers, with diversity being the result of a careless lack of precision. Without a creator the diversity presents no less of a mystery. An undesigned world might well run rampant with eddies and whirls of random fluctuations, yet there is more than enough order in the apparent chaos of life, and just enough elegance in its apparent design, that it stretches credulity to attribute the constitution of biological systems, and the diversity within them, to chance alone.

That said, sometimes variation itself can be rather interesting. The complex web of life, whether deep in the ocean, along the banks of a forest stream or on the peaks of a mountain range, suggests a subtle interdependency. Close examination reveals loops and chains of predators and prey, of life, death, decomposition and new life. And as the landscape changes over time or space, so too do the species that inhabit it. Such variation in landscape also corresponds to variation among very similar species – an observation made famous by Darwin's finches.

It seems clear that some variation is indeed the product of stochastic forces, and some is the result of mishap or misadventure. Yet close scrutiny reveals a picture of life suggestive of deliberate variation. A significant proportion of the diversity observed amongst the myriad arrays of life forms appears to be closely linked to the diversity of environments – both the physical and the social – in which they exist. Thus, at least some variation is not random, but emerges from the forces of evolution. Thus it was the light of evolution that helped illuminate the subtle forces that underpin the phenomena of biological diversity.

17.0.1: Moral diversity

The existence of moral diversity also presents something of a puzzle. Within any group – or even over any dinner table – there is likely to be significant variation in what actions and norms individuals believe to be morally right or wrong, such that no two people are likely to have perfectly identical moral attitudes. The moral disagreement that tends to correspond with moral diversity may even prove intractable.

Yet moral diversity has often been seen as a hurdle to be overcome en route to discovering the correct answers to moral questions. Moral diversity has often been considered as something to explain away rather than something deserving of an explanation. However, this view does little justice to the patterns and regularities that appear in moral diversity. As in biology, moral diversity is sometimes rather interesting, and can hint at deeper forces at work. The complex web of behavioural strategies and norms that influence how we act in a social context, and the environments in which they tend to emerge, is suggestive of a subtle interdependency. As the environment varies, so too do the norms that best solve the problems of social living, particularly those concerning cooperative and coordinated behaviour. Sometimes it even takes multiple behavioural strategies or norms working in concert to best raise cooperation or coordination to high levels with any kind of stability. Thus the light of evolution – both biological and cultural – can also help to reveal some of the subtle forces that underpin the phenomena of moral diversity.

17.1: Moral ecology and diversity

The main lesson to be gleaned from the moral ecology perspective is that moral systems should be expected to be dynamic rather than static. One way to look at morality is as a cultural technology the function of which is to solve the problems of social living in order to facilitate greater levels of prosocial and cooperative behaviour. As such, morality can be seen as a product of social creatures attempting to live together for mutual advantage in a wide range of disparate external and internal environments. The problems of social living are diverse, and themselves change as social dynamics change. As such, there is often no simple way to construct a moral system such as to offer optimal solutions to the problems of social living, and build it in such a way that it is both resistant to invasion by sub-optimal variations yet also amenable to revision by more optimal ones as the environment changes.

Certainly, much of our moral language seems to tell a very different story about morality. If one adopts the inside-out perspective on morality, and takes our moral utterances at face value, it might appear that morality is underpinned by objective facts about which actions or judgments are right or wrong. Or it might appear that morality consists of prescriptive assertions backed by emotive force. Or perhaps some combination of the two. Indeed, there appears to be no single way to interpret our moral language in order to derive some insight into the nature of morality, although efforts to do so continue.

Moral ecology looks at morality from the opposite outside-in point of view. This perspective suggests that, despite what our moral language might imply, morality is created rather than discovered. However, this is not to suggest that "anything goes" in morality. Given a certain environment and the goal of solving the problems of social living, there are likely to be objectively better or worse solutions to those problems in that environment. Yet it remains that no single system of moral norms is likely to effectively solve those problems in every environment. And given the messy process of cultural evolution, we would expect a great deal of variation between cultures when it comes to their solutions to the problems of social living.

Yet this variation in moral systems is not just an artefact of ignorance that we should attempt to explain away. To the extent that environments vary, even if we are entirely comprehending of that variation, there will likely be multiple moral systems that will function well in those environments.

17.2: Ways of life

As discussed in chapter 3, John Mackie famously asked whether moral diversity and disagreement might be better explained in terms of "people's adherence to and participation in different ways of life" rather than the "hypothesis that they express perceptions, most of them seriously inadequate and badly distorted, of objective values" (Mackie, 1977). He suggested that an individual's support for a particular practice – say, monogamy – stems from their monogamous way of life rather than the other way around. However, this raises the question of what influences the various ways of life? Do they vary arbitrarily? And how does variation in ways of life influence variation in moral attitudes?

In a sense, this thesis lends some support to Mackie's conjecture. The dynamics of moral ecology help to explain why ways of life might vary and how variation in ways of life contributes to problems of social living, and ultimately to the diversity of solutions to those problems. Mackie himself subscribed to a kind of moral functionalism. He was aware that many of the problems of social living can be modelled by game theory, and he suggested that morality was created rather than discovered. In these notions he and I are in agreement. In many ways, moral ecology simply adds flesh to the bones of Mackie's picture. I have argued that much of the diversity in moral norms and attitudes that exists among cultures and individuals can be pegged to the variation in the internal and external environments faced by individuals within those cultures. Thus moral ecology emphasises that the various ways of life do not just vary randomly or arbitrarily. Instead, these ways of life are intricately linked to our nature as social animals, and to the dynamics that emerge from social living. This suggests that moral diversity is not just due to ignorance or subjective whim, but is at least in part a reasonable response to varying environmental conditions. While it is not possible to rule out that there do exist objective moral facts, or that sufficiently situated ideally rational individuals might not agree on all moral matters, such notions do little to help understand moral diversity as it exists in the world today. Furthermore, such views risk dismissing much of what is interesting in moral diversity.

17.2.1: Analogies

In chapter 3 I mentioned Paul Bloomfield's analogy between morality and health (Bloomfield, 2001). He suggests that just as there exist facts about health, but these facts can also entertain disagreement over what constitutes a healthy course of action without implying scepticism, so too are there facts about morality, but these facts can entertain a disagreement over what is moral without implying scepticism. There do appear to be some parallels, although I agree with Richard Joyce that the analogy falls apart because moral discourse appears to possess a key feature that health discourse does not: the apparent inescapability of moral prescriptions. One might quite rationally (if imprudently) agree that a certain course of action will lead to good health, yet reject that course of action because they care not one whit for their own health. On the other hand, one cannot simply refuse to follow a moral course of action because they care not one whit for morality (Joyce, 2003).

While Bloomfield intends for his analogy to support moral realism, the very error that undermines his argument actually makes his analogy rather apt for understanding an alternate anti-realist conception of morality – with only a single revision. If it were the case that behaving in an unhealthy manner attracted condemnation and punishment from others, then health discourse would very likely resemble moral discourse. There would remain no firm objective binding commitment to pursue health *no matter what we desire*, but there is a reason why people might still pursue health, such as through a desire to avoid punishment. This is in addition to the contingent desire to reap the benefits of a healthy lifestyle. Perhaps over time individuals will internalise their norms concerning healthy behaviour and the requirement for explicit rules and the threat of punishment might diminish.

Likewise with morality. Even without any binding objective moral facts, individuals may come to behave morally either because they wish to avoid punishment, or because they enjoy the benefits of a more coordinated and cooperative social existence. And if they opt out of morality, those with whom they interact are not likely to entertain their amoralism for long. Note that this does not require that our moral – or health – language make a commitment to the existence of binding objective facts. The discourse could be in error in this regard, or entirely hypothetical, and the end result would be just the same.

Perhaps a better analogy is food preferences. Every culture's cuisine serves the same basic function of providing nutrients in a palatable form. Yet cuisines vary widely across the globe and throughout history. They also vary widely among individuals within any particular culture. The between-culture variation is likely in large part due to environmental variation, particularly in the form of the different availability of various raw materials in the physical environment, as scaffolded by practices like farming and herding. Building on this variation in raw materials is the cultural evolutionary process, whereby new food preparation practices and recipes are innovated and spread through the population. The existence of some ingredients and practices will then facilitate the use of other ingredients; yeast is not terribly useful until the advent of baking and brewing, for example. The development of techniques for preserving foods is another example of a practice that increases our options in terms of ingredients and means of preparation. The end result is a complex web of interdependencies between the environment, available ingredients and cultural knowledge and practices that yields what might be called "cuisine ecology." Variation in cuisine is not indicative of ignorance or error, but a result of different cultures responding to the same requirement to produce palatable nutrients given their environmental contingencies.

There is ample evidence that one's cultural background and their exposure to foods and flavours at a young age strongly influence food preferences later in life (Birch, 1999). Yet individuals also express different tastes, and these influence their food preferences as well. This variation in taste is largely influenced by biological variation, such as in the number and concentration of taste buds an individual has (I. J. Miller & Reedy, 1990), and this variation may indeed be influenced by the same evolutionary forces discussed in chapter 14. The end result is a startlingly diverse array of foods and food preferences. This diversity does not suggest there are no facts about what is nutritious or palatable. But it does suggest there is no "right" answer as to what one ought to eat. In fact, there might be many right answers. We even occasionally employ categorical language about food preferences ("Vegemite is bad, don't eat it!") even though there is no objective basis to ratify such claims.

This is, broadly speaking, a subjectivist account of morality (or health, or cuisine etc.). Or, at least, it is compatible with a subjectivist interpretation, even if it does not necessarily imply that objectivism must be false. However, the subjectivism that underlies this account of morality need not undermine the importance or the prescriptive force of morality. It is likely a simple matter of contingent empirical fact that most people desire to reap the rewards of a social and cooperative existence. And it is likely a simple matter of contingent empirical fact that anyone who opts out of conforming to a system of moral norms will likely be punished into submission by their peers, or perhaps ostracised or exiled. There seems little doubt that most people internalise their culture's system of moral norms and habituate their behaviour (or shape their worldview) to largely, without reflection, behave in accordance with it. But should such habituation or tendency to conform fail to motivate moral behaviour, there typically exists such external motivation, in the form of both positive reinforcement and the threat of punishment, as is required to set them back on the moral track. And as Philippa Foot (1972), Joshua Greene (2002) and Richard Joyce (2006; 2001), among others, have pointed out, the lack of some objective binding commitment to act morally does not necessarily subvert moral behaviour. There seems little likelihood that people will suddenly abandon all interest in living socially or cooperatively, or caring about such things as justice or the welfare of others.

In some ways moral ecology also resembles the moral relativism of Gilbert Harman and David Wong. They both suggest that what is considered right or wrong ought to be assessed relative to a particular moral framework. And these moral frameworks are conventions created by groups of people to facilitate social living (Harman & Thompson, 1996; D. B. Wong, 2006; D. Wong, 1984). However, as I pointed out in chapter 3, these relativist accounts have difficulties in accounting for how an individual with overlapping cultural or group identities ought to reconcile their moral systems if they conflict. Or how an individual with a moral code adopted from one environment ought to adjust their moral views when moving into a different environment. Or how a moral system itself ought to adjust if its environment changes. I would suggest that moral relativism could benefit from the injection of a functionalist definition of morality along with a deeper acknowledgement of the forces of cultural evolution in shaping systems of moral norms in response to environmental variation. This view suggests norms can be indexed to the environmental contingencies that influence how successful they are at satisfying the function of morality. This is not a strictly objective metric, but one by which individuals from two cultures can reconcile their differences over moral norms. Or how an individual might justify their commitment to some norms rather than others in different contexts.

17.3: From moral ecology to moral dynamics

The primary explanandum of this thesis is the phenomenon of moral diversity. The explanans is moral ecology. However, while moral ecology might be a useful tool for helping us to understand how we got where we are, can it tell us anything about where we ought to go?

As mentioned in the opening chapter, there is a long and chequered history of drawing ethical recommendations from evolutionary theory. One potential pitfall for such an effort often goes by the moniker of the "naturalistic fallacy." In recent decades it has been fashionable to declare that one cannot derive an "ought" from an "is," with a nod to David Hume or G. E. Moore, before moving on to address other concerns, either scientific or moral, depending on which side of the is-ought fence one's research happens to land. Such proclamations, particularly common in literature addressing evolution and evolutionary psychology (Walter, 2006), often seek to give a knock-down argument against any attempt to derive any prescriptive recommendations from evolutionary theory.

Yet one can agree with Hume that any series of "is" statements, no matter how exhaustive, does not imply any "ought" statements without consigning evolution – or moral ecology – to irrelevance in ethics. Indeed, Hume's warning only applies to a special kind of "ought," namely those that genuinely do carry binding unconditional force, such as Kant's categorical imperatives. If such oughts existed, it seems that no amount of descriptive facts could fix what those oughts ought to be. However, if one views morality as a cultural technology that serves the function of solving the problems of social living – a function that is open to negotiation – then the "oughts" of morality are more like hypothetical imperatives. There is no “magic force” behind such norms, as Foot put it at one point (Foot, 1972). And there is no normative gulf between descriptive facts and informing such hypothetical imperatives.

The first step in using evolution to elucidate normative morality is to have a discussion about what the function of morality ought to be. In the past it might have been primarily to solve the problems of social living and facilitate prosocial and cooperative behaviour. However, there is nothing to say that we ought to define the function of morality in such terms in the future. We might agree upon a superior formulation, or might decide to explicitly expand the circle of morality to concern more than just a particular in-group, say, expanding it to include animals or other sentient beings. This is a discussion and genuine debate by itself. Assuming the function of morality can be agreed upon, then new norms can be innovated in an attempt to satisfy that function. That is another discussion and debate that can take place.

Moral ecology can inform this discussion, particularly by emphasising the significance of environmental variation and the complex dynamics of social and cooperative interaction. If, for example, we seek a norm that is effective at coordinating behaviour or distributing resources or managing a hierarchy, then the optimal solution (or solutions) is likely to depend on the state of the environment, both external and internal. If we wish to assess whether an existing norm is effective at satisfying the function of morality, then we can judge it by how it performs in the environment in which it operates.

17.3.1: Moral dynamics

As such, I would suggest that moral ecology could lead to a new discipline, perhaps called “moral dynamics,” which employs the tools of evolution and game theory to inform moral enquiry. Once the function of morality is agreed upon, then moral dynamics can be a useful tool to examine the dynamics of the problem background facing morality given the particular environment in which it is set to operate. The goal would be to move on from the often haphazard process of cultural evolution as a method of advancing moral progress, and instead employ a more deliberate and direct approach to the innovation of new moral norms. It could also emphasise the difficulty of introducing a new norm into an existing moral system and predicting what its outcome is likely to be. Due to the interdependence of moral norms, it could encourage a kind of directed experimentation and steady reform, whereby new norms are implemented, monitored, assessed and refined.

Moral ecology and moral dynamics could also help resolve many moral disagreements by emphasising the importance of diversity in moral norms in response to diversity in environmental conditions. It could also emphasise that many norms are functionally equivalent, thus dissolving some of the disputes that exist between adherents of various moral systems today.

Ultimately, moral ecology brings morality into step with the other natural and social sciences. The prevailing tendency in the Western philosophical tradition to see morality as some supernatural or non-natural construct gives morality a somewhat rarefied quality detached from the messy contingencies of our biological existence. Like with so many things, the introduction of evolution into ethical thinking can bring our gaze firmly down to earth. We are, after all, social animals, and our success as such is largely due to our social nature. But understanding the complexities and dynamics of the social technology we have invented to facilitate our social living can not only make morality scrutable, but also make it relevant to our lives. To appropriate Dobzhansky once again, and not without a whiff of rhetoric, much of ethics can only make sense in the light of evolution.

Bibliography

Abarbanell, L., & Hauser, M. D. (2010). Mayan Morality: An Exploration of Permissible Harms. *Cognition*, *115*(2), 207–24. doi:10.1016/j.cognition.2009.12.007

Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., & Sanford, R. N. (1950). *The Authoritarian Personality*. Oxford: Harper.

Alexander, R. D. (1987). *The Biology of Moral Systems*. New York: Aldine De Gruyter.

Alford, J. R., Funk, C. L., & Hibbing, J. R. (2005). Are Political Orientations Genetically Transmitted? *American Political Science Review*, *99*(02), 153–167.

Alford, J. R., Funk, C. L., & Hibbing, J. R. (2008). Beyond liberals and conservatives to political genotypes and phenotypes. *Perspectives on Politics*, *6*(02), 321–328.

Altemeyer, B. (1981). *Right-wing Authoritarianism*. Winnipeg: University of Manitoba Press.

Altemeyer, B. (1998). The Other Authoritarian Personality. *Advances in Experimental Social Psychology*, *30*, 47–92.

Altemeyer, B. (2003). What Happens When Authoritarians Inherit the Earth? A Simulation. *Analyses of Social Issues and Public Policy*, *3*(1), 15–23.

Altemeyer, B. (2004). Highly dominating, highly authoritarian personalities. *The Journal of Social Psychology*, *144*(4), 421–448.

Altemeyer, B. (2006). *The Authoritarians*. Winnipeg: University of Manitoba Press.

Anson, J., Pyszczynski, T., Solomon, S., & Greenberg, J. (2009). Political Ideology in the 21st Century: A Terror Management Perspective on Maintenance and Change of the Status Quo. In J. T. Jost, A. C. Kay, & H. Thorisdottir (Eds.), *Social and Psychological Bases of Ideology and System Justification* (Vol. 1, pp. 210–241). Oxford, UK: Oxford University Press.

Asch, S. E. (1956). Studies of Independence and Conformity: I. A Minority of One Against a Unanimous Majority. *Psychological Monographs: General and Applied*, *70*(9).

Axelrod, R. (1984). *The Evolution of Cooperation*. New York: Basic Books.

Axelrod, R. (1986). An evolutionary approach to norms. *The American Political Science Review*, *80*(4), 1095–1111.

Axelrod, R. (1987). The Evolution of Strategies in the Iterated Prisoner's Dilemma. In L. Davis (Ed.), *Genetic Algorithms and Simulated Annealing* (pp. 32–41). London: Pitman.

Axelrod, R. (1997). *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. New Jersey: Princeton University Press.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390.

Ayer, A. J. (1936). *Language, Truth and Logic* (Penguin 19.). London: Victor Gollancz.

Baron, J. (1994). Nonconsequentialist decisions. *Behavioral and Brain Sciences*.

Bateson, P. (2004). The origins of human differences. *Daedalus*, *133*(4), 36–46.

Beaumont, H. J. E., Gallie, J., & Kost, C. (2009). Experimental evolution of bet hedging. *Nature*.

Bendor, J., & Swistak, P. (2001). The Evolution of Norms. *American Journal of Sociology*, *106*(6), 1493.

Benyamin, B., Pourcain, B., Davis, O. S., Davies, G., Hansell, N. K., Brion, M.-J., … Visscher, P. M. (2013). Childhood intelligence is heritable, highly polygenic and associated with FNBP1L. *Molecular Psychiatry*. doi:10.1038/mp.2012.184

Berger, P. L., & Luckmann, T. (1966). *The Social Construction of Reality*. Garden City: Anchor Books.

Bergstrom, C. T., & Godfrey-Smith, P. (1998). On the Evolution of Behavioral Heterogeneity in Individuals and Populations. *Biology and Philosophy*, *13*, 205–231.

Bicchieri, C., Duffy, J., & Tolle, G. (2004). Trust Among Strangers. *Philosophy of Science*, *71*(3), 286–319. doi:10.1086/381411

Bingham, P. M. (2000). Human evolution and human history: A complete theory. *Evolutionary Anthropology: Issues, News, and Reviews*, *9*(6), 248–257. doi:10.1002/1520-6505(2000)9:6<248::AID-EVAN1003>3.0.CO;2-X

Binmore, K. (2007). *Playing for Real: A Text on Game Theory*. London: Oxford University Press.

Birch, L. (1999). Development of food preferences. *Annual Review of Nutrition*.

Birkhead, T. R., & Monaghan, P. (2010). Ingenious Ideas: The History of Behavioral Ecology. In D. F. Westneat & C. W. Fox (Eds.), *Evolutionary Behavioral Ecology*. Oxford: Oxford University Press.

Blackburn, S. (1984). *Spreading the Word*. Oxford: Clarendon Press.

Blackburn, S. (1993). *Essays in Quasi-Realism*. New York: Oxford University Press.

Bloomfield, P. (2001). *Moral reality*. Oxford: Oxford University Press.

Bloomfield, P. (2008). Disagreement about Disagreement. In W. Sinnott-Armstrong (Ed.), *Moral Psychology vol. 2*. Cambridge: MIT Press.

Boag, P. T., & Grant, P. R. (1981). Intense Natural Selection in a Population of Darwin's Finches (Geospizinae) in the Galapagos. *Science*, *214*(4516), 82–85.

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

Bouchard, T. J. (1994). Genes, Environment, and Personality. *Science*, *264*, 1700–1701.

Bouchard, T. J., & McGue, M. (2003). Genetic and environmental influences on human psychological differences. *Journal of Neurobiology*, *54*(1), 4–45. doi:10.1002/neu.10160

Bowles, S. (2006). Group competition, reproductive leveling, and the evolution of human altruism. *Science*, *314*(5805), 1569–72. doi:10.1126/science.1134829

Bowles, S. (2008). Being human: Conflict: Altruism's midwife. *Nature*, *456*(7220), 326–7. doi:10.1038/456326a

Bowles, S. (2009). Did warfare among ancestral hunter-gatherers affect the evolution of human social behaviors? *Science*, *324*(5932), 1293–8. doi:10.1126/science.1168112

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *PNAS*, *100*(6), 3531–3535. doi:10.1073/pnas.0630443100

Boyd, R., & Lorberbaum, J. P. (1987). No Pure Strategy is Evolutionarily Stable in the Repeated Prisoner's Dilemma Game. *Nature*, *327*.

Boyd, R. N. (1988). How to be a Moral Realist. In G. Sayre-McCord (Ed.), *Essays on Moral Realism* (pp. 181–228). Ithaca and London: Cornell University Press.

Boyd, R., & Richerson, P. J. (1985). *Culture and the Evolutionary Process* (p. 301). Chicago: University of Chicago Press.

Boyd, R., & Richerson, P. J. (1992). Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizable Groups. *Ethology and Sociobiology*.

Boyd, R., & Richerson, P. J. (1995). Why does culture increase human adaptability? *Ethology and Sociobiology*, *16*, 125–143.

Boyd, R., Richerson, P. J., & Henrich, J. (2011). The cultural niche: why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences of the United States of America*, *108 Suppl*, 10918–25. doi:10.1073/pnas.1100290108

Braendle, C., Davis, G., Brisson, J., & Stern, D. (2006). Wing dimorphism in aphids. *Heredity*.

Brink, D. O. (1984). Moral realism and the sceptical arguments from disagreement and queerness. *Australasian Journal of Philosophy*, *62*(2), 111–125. doi:10.1080/00048408412341311

Brink, D. O. (1989). *Moral Realism and the Foundations of Ethics*. Cambridge: Cambridge University Press.

Buckholtz, J., & Marois, R. (2012). The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. *Nature Neuroscience*, 1–7.

Buss, D. M. (1999). *Evolutionary Psychology: The New Science of the Mind* (4th Editio.). London: Pearson.

Buss, D. M. (2009). How Can Evolutionary Psychology Successfully Explain Personality and Individual Differences? *Perspectives on Psychological Science*, *4*(4), 359.

Byrne, R. W. (1996). Machiavellian intelligence. *Evolutionary Anthropology*, *5*(5), 172–180. doi:10.1002/(SICI)1520-6505(1996)5:5<172::AID-EVAN6>3.0.CO;2-H

Byrne, R. W., & Whiten, A. (1989). *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans*. Oxford: Oxford Science Publications.

Casebeer, W. D., & Churchland, P. S. (2003). The Neural Mechanisms of Moral Cognition: A Multiple-Aspect Approach toMoral Judgment and Decision-Making. *Biology and Philosophy*, *18*(1), 169–194.

Chalmers, D. J. (1996). *The Conscious Mind*. Oxford: Oxford University Press.

Charlton, B. (1997). Injustice, Inequality and Evolutionary Psychology'. *Journal of Health Psychology*, *2*, 413–25.

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology*, *9*(2), 129–36.

Childs, D. Z., Metcalf, C. J. E., & Rees, M. (2010). Evolutionary bet-hedging in the real world: empirical evidence and challenges revealed by plants. *Proceedings. Biological Sciences / The Royal Society*, *277*(1697), 3055–64. doi:10.1098/rspb.2010.0707

Clark, A. (2008). *Supersizing the Mind*. Oxford: Oxford University Press.

Coleman, J. (1998). *Foundations of Social Theory*. Cambridge: Harvard University Press.

Cook, L. M. (2003). The Rise and Fall of the Carbonaria Form of the Peppered Moth. *The Quarterly Review of Biology*, *78*(4), 399–417.

Cosmides, L., & Tooby, J. (1997). Evolutionary psychology: A primer.

Cosmides, L., & Tooby, J. (2004). Knowing thyself: The evolutionary psychology of moral reasoning and moral sentiments. *Business, Science, and Ethics*, *4*, 91–127.

Cummins, R. (1975). Functional Analysis. *The Journal of Philosophy*, *72*(20), 741. doi:10.2307/2024640

Cummins, R. (2002). Neo-Teleology. In A. Ariew, R. Cummins, & M. Perlman (Eds.), *Functions: New Essays in the Philosophy of Psychology and Biology* (pp. 157–172). Oxford: Oxford University Press.

Curtis, V., de Barra, M., & Aunger, R. (2011). Disgust as an adaptive system for disease avoidance behaviour. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *366*(1563), 389–401. doi:10.1098/rstb.2010.0117

Cushman, F., Knobe, J., & Sinnott-Armstrong, W. (2008). Moral appraisals affect doing/allowing judgments. *Cognition*.

Cushman, F., Young, L., & Hauser, M. D. (2006). The role of conscious reasoning and intuition in moral judgment testing three principles of harm. *Psychological Science*.

d'Errico, F., Henshilwood, C., Vanhaeren, M., & van Niekerk, K. (2005). Nassarius kraussianus shell beads from Blombos Cave: evidence for symbolic behaviour in the Middle Stone Age. *Journal of Human Evolution*, *48*(1), 3–24. doi:10.1016/j.jhevol.2004.09.002

Dall, S. R. X., Houston, A., & McNamara, J. M. (2004). The behavioural ecology of personality: consistent individual differences from an adaptive perspective. *Ecology Letters*.

Darwin, C. (1872). *On the Origin of Species* (6th editio.). London: Project Gutenberg.

Davis, B., & Dossetor, K. (2010). (Mis) perceptions of crime in Australia. *Trends and Issues in Crime and ...*.

Dawkins, R. (1982). *The Extended Phenotype*. Oxford: Oxford University Press.

Dawkins, R. (2006). *The Selfish Gene - 30th Anniversary Edition* (3rd ed.). New York: Oxford University Press.

De Moor, M. H. M., Costa Jr., P. T., Terracciano, A., Krueger, R. F., De Geus, E. J., Toshiko, T., … Boomsma, D. I. (2010). Meta-analysis of genome-wide association studies for personality. *Molecular Psychiatry*, (April), 1–13. doi:10.1038/mp.2010.128

De Waal, F. (1982). *Chimpanzee Politics: Sex and Power Among Apes. London, UK: Jonathan Cape* (25th Anniv.). New York: Harper & Row.

De Waal, F. (2006). *Primates and Philosophers*. Princeton: Princeton University Press.

Dean, T. (2012). Evolution and Moral Diversity. *The Baltic International Yearbook of Cognition, Logic and Communication*, *7*(October), 1–16. doi:10.4148/biyclc.v7i0.1775

DeFranco, A. L., Locksley, R. M., & Robertson, M. (2007). The MHC and Polymorphism of MHC Molecules. In *Immunity: The Immune Response in Infectious and Inflammatory Disease*. Sunderland: New Science Press.

Dewey, J. (1922). *Human Nature and Conduct*. New York: Henry Holt.

Diamond, J. (2012). *The World Until Yesterday* (p. 499). London: Penguin Books.

Dobzhansky, T. (1973). Nothing in biology makes sense except in the light of evolution. *The American Biology Teacher*, *68*(1).

Doris, J. M., & Plakias, A. (2008). How to Argue about Disagreement: Evaluative Diversity and Moral Realism. In W. Sinnott-Armstrong (Ed.), *Moral Psychology vol. 2*. Cambridge, MA: MIT Press.

Dubreuil, B., & Grégoire, J.-F. (2012). Are moral norms distinct from social norms? A critical assessment of Jon Elster and Cristina Bicchieri. *Theory and Decision*, *75*(1), 137–152. doi:10.1007/s11238-012-9342-3

Duckitt, J. (2001). A dual-process cognitive-motivational theory of ideology and prejudice. *Advances in Experimental Social Psychology*, *33*, 41–113.

Duckitt, J., & Sibley, C. G. (2009). A Dual Process Motivational Model of Ideological Attitudes and System Justification. In J. T. Jost, A. C. Kay, & H. Thorisdottir (Eds.), *Social and Psychological Bases of Ideology and System Justification* (p. 292).

Dunbar, R. I. M. (1992). Neocortex size as a constraint size in primates on group ecologically. *Journal of Human Evolution*, *20*, 469–493.

Dunbar, R. I. M. (1998). *Grooming, Gossip, and the Evolution of Language*. Cambridge: Harvard University Press.

Dunbar, R. I. M. (2003a). The Social Brain: Mind, Language, and Society in Evolutionary Perspective. *Annual Review of Anthropology*, *32*(1), 163–181. doi:10.1146/annurev.anthro.32.061002.093158

Dunbar, R. I. M. (2003b). Why are apes so smart. *Primate Life Histories and Socioecology*.

Durkheim, E. (1895). *The Rules of Sociological Method*. (W. D. Halls, Trans.) (Free Press.). New York: Simon & Schuster.

Durkheim, E. (1915). *The Elementary Forms of Religious Life*. (J. W. Swain, Ed.) (Dover.). London: George Allen & Unwin.

Dwyer, S. (2009). Moral dumbfounding and the linguistic analogy: Methodological implications for the study of moral judgment. *Mind & Language*, *24*(3), 274–296.

Earle, A. M. (1896). *Curious Punishments Of Bygone Days*. Chicago: Herbert S. Stone.

Earley, R. L., & Dugatkin, L. A. (2010). Behavior in Groups. In D. F. Westneat & C. W. Fox (Eds.), *Evolutionary Behavioral Ecology*. Oxford: Oxford University Press.

Emery, N. J., Clayton, N. S., & Frith, C. D. (2007). Introduction. Social intelligence: from brain to culture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 485–488. doi:10.1098/rstb.2006.2022

Erikson, R. S., & Tedin, K. L. (2011). *American Public Opinion: Its Origins, Content, and Impact* (8th Editio., p. 416). New Jersey: Pearson.

Fehr, E. (2004). Don't lose your reputation. *Nature*, *432*(7016), 449–450.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–40. doi:10.1038/415137a

Feldman, M. W., & Cavalli-Sforza, L. L. (1985). *On the Theory of Evolution Under Genetic and Cultural Transmission with Application to the Lactose Absorption Problem*. Stanford: Stanford Institute for Population and Resource Studies.

Finlayson, C. (2009). *The Humans Who Went Extinct: Why Neanderthals Died Out and We Survived*. Oxford: Oxford University Press.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection* (p. 318). Oxford: Oxford University Press.

Flanagan, O., Sarkissian, H., & Wong, D. (2008). Naturalizing Ethics. In W. Sinnott-Armstrong (Ed.), *Moral Psychology* (pp. 1–31). Cambridge, MA: MIT Press.

Foot, P. (1972). Morality as a system of hypothetical imperatives. *The Philosophical Review*, *81*(3), 305–316.

Fowler, J. H. (2005). Altruistic Punishment and the Origin of Cooperation. *Proceedings of the National Academy of Sciences*.

Fowler, J. H. (2006). Altruism and Turnout. *Journal of Politics*.

Fraser, B., & Hauser, M. D. (2010). The Argument from Disagreement and the Role of Cross- Cultural Empirical Data. *Mind & Language*, *8329*.

Frenkel-Brunswik, E. (1948). Tolerance toward ambiguity as a personality variable. *American Psychologist*, *3*, 268.

Gauthier, D. (1986). *Morals by Agreement*. Oxford: Oxford University Press.

Geach, P. T. (1965). Assertion. *The Philosophical Review1*, *74*(4), 449–465.

Ghalambor, C. K., Angeloni, L. M., & Carroll, S. P. (2010). Behavior as Phenotypic Plasticity. In D. F. Westneat & C. W. Fox (Eds.), *Evolutionary Behavioral Ecology*. Oxford: Oxford University Press.

Ghiselin, M. T. (1974). *The economy of nature and the evolution of sex*. Berkeley: University of California Press (Berkeley).

Gigerenzer, G. (2008). Moral Intuition = Fast and Frugal Heuristics? In W. Sinnott-Armstrong (Ed.), *Moral Psychology vol. 2*. Cambridge: MIT Press.

Gil-White, F. J., & Richerson, P. J. (2002). Large scale human cooperation and conflict. In *Encyclopedia of Cognitive Science*.

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior*, *24*(3), 153–172.

Godfrey-Smith, P. (1994). A modern history theory of functions. *Nous*, *28*(3), 344–362.

Godfrey-Smith, P. (1998). *Complexity and the Function of Mind in Nature* (p. 328). Cambridge: Cambridge University Press.

Gopnik, A. (2009). *The Philosophical Baby*. London: The Bodley Head.

Gould, S., & Vrba, E. (1982). Exaptation-a missing term in the science of form. *Paleobiology*.

Graham, J., Nosek, B. a, Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366–85. doi:10.1037/a0021847

Grant, B. R., & Grant, P. R. (1989). *Evolutionary Dynamics of a Natural Population: The Large Cactus Finch of the Galápagos*. Chicago: University of Chicago Press.

Grant, B. S. (1999). Fine Tuning the Peppered Moth Paradigm. *Evolution*, *53*(3), 980–984.

Grant, P. R. (1986). *Ecology and Evolution of Darwin's Finches* (2nd editio.). Princeton: Princeton University Press.

Grant, P. R., & Grant, B. R. (2002). Unpredictable Evolution in a 30-Year Study of Darwin's Finches. *Science*, *296*(5568), 707–711.

Grant, P. R., & Grant, B. R. (2006). Evolution of Character Displacement in Darwin's Finches. *Science*, *313*(5784), 224–6. doi:10.1126/science.1128374

Greene, J. (2002). *The terrible, horrible, no good, very bad truth about morality and what to do about it*. Princeton University.

Greene, J., Cushman, F., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: the interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364–71. doi:10.1016/j.cognition.2009.02.001

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, *6*(12), 517–523.

Greene, J., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–8. doi:10.1126/science.1062872

Griffiths, P. E. (1993). Functional analysis and proper functions. *The British Journal for the Philosophy of Science*, *44*(3), 409.

Gross, M. R. (1982). Sneakers, Satellites and Parentals: Polymorphic Mating Strategies in North American Sunfishes. *Zeitschrift Für Tierpsychologie*, *60*(1).

Gross, M. R. (1996). Alternative reproductive strategies and tactics: diversity within sexes. *Trends in Ecology & Evolution*, *11*(2), 92–98.

Guhl, A. M., Collias, N. E., & Allee, W. C. (1945). Mating Behavior and the Social Hierarchy in Small Flocks of White Leghorns. *Physiological Zoology*, *18*(4), 365–390.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834.

Haidt, J. (2003). The Moral Emotions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of Affective Sciences* (pp. 852–870). Oxford: Oxford University Press.

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, *316*(5827), 998–1002. doi:10.1126/science.1137651

Haidt, J., Björklund, F., & Murphy, S. (2000). *Moral Dumbfounding: When Intuition Finds No Reason*.

Haidt, J., & Graham, J. (2007). When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals may not Recognize. *Social Justice Research*, *20*(1), 98–116.

Haidt, J., & Graham, J. (2009). Planet of the Durkheimians, Where Community, Authority, and Sacredness Are Foundations of Morality. In J. T. Jost, A. C. Kay, & H. Thorisdottir (Eds.), *Social and Psychological Bases of Ideology and System Justifi cation* (pp. 371–401). Oxford, UK: Oxford University Press.

Haidt, J., & Hersh, M. A. (2001). Sexual Morality: The Cultures and Emotions of Conservatives and Liberals. *Journal of Applied Social Psychology*, *31*(1), 191–221.

Haidt, J., & Kesebir, S. (2010). Morality. In S. Fiske, D. Gilbert, & G. Lindzey (Eds.), *The Handbook of Social Psychology* (5th Editio., Vol. 53, pp. 797–832). Wiley.

Haidt, J., Roller, S. H., & Dias, M. G. (1993). Affect, Culture, and Morality, or Is It Wrong to Eat \bur Dog? *Journal of Personality and Social Psychology*, *65*(4), 613–623.

Haidt, J., Rosenberg, E., & Hom, H. (2001). Differentiating diversities: Moral diversity is not like other kinds. *Journal of Applied Social Psychology*, 1–38.

Haley, K., & Fessler, D. M. T. (2005). Nobody's watching?: Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior*.

Hamilton, W. D. (1963). The Evolution of Altruistic Behavior. *The American Naturalist*, *97*(896), 354–356.

Hansell, M. H. (1984). *Animal Architecture and Building Behaviour. Animal architecture and building behaviour.* London: Longman.

Hardin, G. (1968). The Tragedy of the Commons. *Science*, *162*(3859), 1243–8. doi:10.1126/science.162.3859.1243

Hare, R. D. (1999). *Without Conscience: The Disturbing World of the Psychopaths Among Us*. New York: Guilford Press.

Harman, G. (1991). Moral Diversity as an Argument for Moral Relativism. In D. Odegard & C. Stewart (Eds.), *Perspectives on Moral Relativism*. Milliken: Agathon Books.

Harman, G., & Thompson, J. (1996). *Moral Relativism and Moral Objectivity*. Cambridge, MA: Blackwell Publishers.

Harvell, C. D. (1998). Genetic variation and polymorphism in the inducible spines of a marine bryozoan. *Evolution*.

Haselton, M., & Buss, D. M. (2002). Biases in Social Judgment: Design Flaws or Design Features? In J. Forgas, W. von Hippel, & K. Williams (Eds.), *Responding to the Social World: Explicit and Implicit Processes in Social Judgments and Decisions*. Psychology Press.

Hauser, M. D. (2006). *Moral Minds*. New York: HarperCollins.

Hawkes, K., O'Connel, J. F., Blurton Jones, N. G., Alvarez, H., & Charnov, E. L. (1998). Grandmothering, menopause, and the evolution of human life histories. *Proceedings of the National Academy of Sciences*, *95*(3), 1336–1339.

Hechter, M., & Kanazawa, S. (1997). Sociological Rational Choice Theory. *Annual Review of Sociology*, *23*.

Henrich, J. (2009). The evolution of costly displays, cooperation and religion. *Evolution and Human Behavior*, *30*(4), 244–260. doi:10.1016/j.evolhumbehav.2009.03.005

Henrich, J., & Boyd, R. (1998). The Evolution of Conformist Transmission and the Emergence of Between-Group Differences. *Evolution and Human Behavior*, *19*, 215–241.

Henrich, J., Boyd, R., & Richerson, P. J. (2012). The puzzle of monogamous marriage. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1589), 657–669. doi:10.1098/rstb.2011.0290

Henrich, J., & Henrich, N. (2010). The evolution of cultural adaptations: Fijian food taboos protect against dangerous marine toxins. *Proceedings. Biological Sciences / The Royal Society*, *277*(1701), 3715–24. doi:10.1098/rspb.2010.1191

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., … Ziker, J. (2006). Costly punishment across human societies. *Science*, *312*(5781), 1767–70. doi:10.1126/science.1127333

Herodotus. (1996). *The Histories*. London: Penguin Books.

Huebner, B., Dwyer, S., & Hauser, M. D. (2009). The role of emotion in moral psychology. *Trends in Cognitive Sciences*, *13*(1), 1–6. doi:10.1016/j.tics.2008.09.006

Hume, D. (1739). *A Treatise of Human Nature* (Penguin.). London: Penguin Books.

Humphrey, N. K. (1976). The social function of intellect. *Growing Points in Ethology*.

Jackson, F., & Pettit, P. (1995). Moral functionalism and moral motivation. *The Philosophical Quarterly*.

Jost, J. T. (2006). The end of the end of ideology. *The American Psychologist*, *61*(7), 651–70. doi:10.1037/0003-066X.61.7.651

Jost, J. T., Federico, C. M., & Napier, J. L. (2009). Political ideology: its structure, functions, and elective affinities. *Annual Review of Psychology*, *60*, 307–37. doi:10.1146/annurev.psych.60.110707.163600

Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Political conservatism as motivated social cognition. *Psychological Bulletin*, *129*(3), 339–375.

Jost, J. T., Kay, A. C., & Thorisdottir, H. (2009). *Social and Psychological Bases of Ideology and System Justification*. (J. T. Jost, A. C. Kay, & H. Thorisdottir, Eds.). Oxford, UK: Oxford University Press.

Jost, J. T., Napier, J. L., Thorisdottir, H., Gosling, S. D., Palfai, T. P., & Ostafin, B. (2007). Are needs to manage uncertainty and threat associated with political conservatism or ideological extremity? *Personality and Social Psychology Bulletin*, *33*(7), 989–1007. doi:10.1177/0146167207301028

Joyce, R. (2001). *The Myth of Morality*. Cambridge: Cambridge University Press.

Joyce, R. (2003). Moral Reality review. *Mind*, *112*.

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

Kant, I. (1785). *Groundwork of the Metaphysic of Morals*. (H. J. Paton, Trans.) (Reprinted .). New York: Harper & Row.

Keller, M. C., & Miller, G. F. (2006). Resolving the paradox of common, harmful, heritable mental disorders: which evolutionary genetic models work best? *Behavioral and Brain Sciences*, *29*(4), 385–404. doi:10.1017/S0140525X06009095

Kelly, D., Stich, S., Haley, K. J., Eng, S. J., & Fessler, D. M. T. (2007). Harm, Affect, and the Moral/Conventional Distinction. *Mind & Language*, *22*(2), 117–131. doi:10.1111/j.1468-0017.2007.00302.x

Kerlinger, F. N. (1984). *Liberalism and Conservatism: The Nature and Structure of Social Attitudes. Beyond Liberal and Conservative: Reassessing the Political Spectrum*. Hillsdale: Erblaum.

Kettlewell, B. (1973). *The Evolution of Melanism*. Oxford: Clarendon Press.

Kitcher, P. (1993). Function and Design. *Midwest Studies in Philosophy*, *18*(1), 379–397. doi:10.1111/j.1475-4975.1993.tb00274.x

Kitcher, P. (1998). Psychological Altruism, Evolutionary Origins, and Moral Rules. *Philosophical Studies*, *89*, 283–316.

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

Kohlberg, L., & Hersh, R. H. (1977). Moral development: A review of the theory. *Theory into Practice*, *16*(2), 53–59.

Krebs, J. R., & Davies, N. B. (1978). The Evolution of Behavioural Ecology. In J. R. Krebs & N. B. Davies (Eds.), *Behavioural Ecology: An Evolutionary Approach* (4th ed., pp. 3–12). London: Blackwell Science.

Lack, D. (1961). *Darwin's Finches* (2nd editio.). New York: Harper.

Lakoff, G. (1996). *Moral Politics*. Chicago: University of Chicago Press.

Lee, K. E. (1985). *Earthworms: Their Ecology and Relationships with Soils and Land Use*. Waltham: Academic Press.

Leiter, B. (2008). Against Convergent Moral Realism: The Respective Roles of Philosophical Argument and Empirical Evidence. In W. Sinnott-Armstrong (Ed.), *Moral Psychology vol. 2*. Cambridge: MIT Press.

Levins, R. (1968). *Evolution in Changing Environments*. Princeton University Press.

Lewontin, R. C. (1983). Gene, Organism and Environment. In D. S. Bendall (Ed.), *Evolution from Molecules to Men*. Cambridge: Cambridge University Press.

Lieberman, D. (2008). Moral Sentiments Relating to Incest: Discerning Adaptations from By-products. In W. Sinnott-Armstrong (Ed.), *Moral Psychology vol. 1* (pp. 165–190). Cambridge, MA: The MIT Press.

Lieberman, D., Tooby, J., & Cosmides, L. (2003). Does morality have a biological basis? An empirical test of the factors governing moral sentiments relating to incest. *Proceedings. Biological Sciences / The Royal Society*, *270*(1517), 819–26. doi:10.1098/rspb.2002.2290

Loeb, D. (2008). Moral Incoherentism: How to Pull a Metaphysical Rabbit out of a Semantic Hat. In W. Sinnott-Armstrong (Ed.), *Moral Psychology vol. 2*. Cambridge, MA: MIT Press.

Lomborg, B. (1996, April). Nucleus and Shield: The Evolution of Social Structure in the Iterated Prisoner's Dilemma. *American Sociological ReviewReview*, p. 278. doi:10.2307/2096335

MacDonald, K. B. (1998). Evolution, culture, and the five-factor model. *Journal of Cross-Cultural Psychology*, *29*(1), 119.

Mackie, J. L. (1977). *Ethics: Inventing Right and Wrong*. London: Penguin Books.

Martin, J. L. (2001). The Authoritarian Personality, 50 Years Later: What Questions Are There for Political Psychology? *Political Psychology*, *22*(1), 1–26.

Maynard Smith, J. (1964). Group Selection and Kin Selection. *Nature*, *201*(4924), 1145–1147. doi:10.1038/2011145a0

Maynard Smith, J. (1982). The Hawk-Dove Game. In *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.

Maynard Smith, J. (1986). Evolutionary Game Theory. *Physica*, 43–49.

Maynard Smith, J., & Price, G. R. (1973). The Logic of Animal Conflict. *Nature*, *246*(5427), 15–18.

Mayr, E. (1961). Cause and Effect in Biology. *Science*, *134*(3489), 1501–1506. doi:10.1126/science.134.3489.1501

McCoy S, & Major, B. (2007). Priming meritocracy and the psychological justification of inequality. *Journal of Experimental Social Psychology*, *43*(3), 341–351. doi:10.1016/j.jesp.2006.04.009

McCrae, R. R., & Costa Jr., P. T. (2003). *Personality in Adulthood: A Five-factor Theory Perspective*. New York: The Guilford Press.

McCrae, R. R., & Costa Jr., P. T. (2008). The Five-Factor Theory of Personality. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of Personality*. New York: The Guilford Press.

McElreath, R., Boyd, R., & Richerson, P. J. (2003). Shared Norms and the Evolution of Ethnic Markers. *Current Anthropology*, *44*(1), 122–130. doi:10.1086/345689

McGill, B. J., & Brown, J. S. (2007). Evolutionary Game Theory and Adaptive Dynamics of Continuous Traits. *Annual Review of Ecology, Evolution, and Systematics*, *38*(1), 403–435. doi:10.1146/annurev.ecolsys.36.091704.175517

Mealy, L. (1997). The sociobiology of sociopathy: An integrated evolutionary model. In S. Baron-Cohen (Ed.), *The Maladapted Mind: Classic Readings in Evolutionary Psychopathology*. Hove: Psychology Press.

Merton, R. K. (1968). *Social theory and social structure* (1968 enlar.). New York: Free Press.

Meyers, L. A., & Bull, J. J. (2002). Fighting change with change: adaptive variation in an uncertain world. *Trends in Ecology & Evolution*, *17*(12), 551–557. doi:10.1016/S0169-5347(02)02633-2

Milgram, S. (1974). *Obedience to Authority: An Experimental View*. London: Tavistock Publications.

Mill, J. S. (1859). *On Liberty*. New York: Barnes & Noble.

Miller, G. F. (2001). *The Mating Mind: How sexual choice shaped the evolution of human nature*. New York: Anchor Books.

Miller, I. J., & Reedy, F. E. (1990). Variations in human taste bud density and taste intensity perception. *Physiology & Behavior*, *47*(6), 1213–1219. doi:10.1016/0031-9384(90)90374-D

Millikan, R. G. (1989). In Defense of Proper Functions. *Philosophy of Science*, *56*(2), 288–302.

Mitchell, G., & Tetlock, P. E. (2009). Disentangling Reasons and Rationalizations: Exploring Perceived Fairness in Hypothetical Societies. In J. T. Jost, A. C. Kay, & H. Thorisdottir (Eds.), *Social and Psychological Bases of Ideology and System Justification* (Vol. 1, pp. 126–158). Oxford, UK: Oxford University Press.

Molander, P. (1985). The optimal level of generosity in a selfish, uncertain environment. *Journal of Conflict Resolution*.

Mondak, J. J. (2010). *Personality and the Foundations of Political Behavior*. Cambridge: Cambridge University Press.

Moody-Adams, M. M. (1997). *Fieldwork in Familiar Places: Morality, Culture, and Philosophy*. Cambridge: Harvard University Press.

Moore, G. E. (1903). *Principia Ethica*. New York: Barnes & Noble, 2005.

Mootha, V., & Hirschhorn, J. (2010). Inborn variation in metabolism. *Nature Genetics*.

Mulder, E. J., van Baal, C., Gaist, D., Kallela, M., Kaprio, J., Svensson, D. A., … Palotie, A. (2003). Genetic and Environmental Influences on Migraine: A Twin Study Across Six Countries. *Twin Research*, *6*(5), 442–431.

Naiman, R. J., Johnston, C. A., & Kelley, J. C. (1988). Alteration of North American Streams by Beaver. *BioScience*, *38*(11).

Neander, K. (1991). The teleological notion of "function." *Australasian Journal of Philosophy*, *69*(4), 454–468. doi:10.1080/00048409112344881

Nettle, D. (2005). An evolutionary approach to the extraversion continuum. *Evolution and Human Behavior*, *26*(4), 363–373. doi:10.1016/j.evolhumbehav.2004.12.004

Nettle, D. (2006). The evolution of personality variation in humans and other animals. *American Psychologist*.

Nisbett, R. E., & Cohen, D. (1996). *Culture of honor: The psychology of violence in the South.* (p. 119). Boulder, Colorado: Westview Press.

Norenzayan, A. (2010). Why We Believe: Religion as a Human Universal. In H. Høgh-Olesen (Ed.), *Human Morality and Sociality*. Basingstoke: Palgrave Macmillan.

Nowak, M., & Highfield, R. (2011). *Super Cooperators: Evolution, altruism and human behaviour (or why we need each other to succeed)*. Melbourne: Penguin Books.

Nowak, M., & May, R. M. (1992). Evolutionary games and spatial chaos. *Nature*, *359*(6398), 826–829.

Nowak, M., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, *437*(7063), 1291–8. doi:10.1038/nature04131

O'Brien, D. T., & Wilson, D. S. (2011). Community perception: the ability to assess the safety of unfamiliar neighborhoods and respond adaptively. *Journal of Personality and Social Psychology*, *100*(4), 606–20. doi:10.1037/a0022803

Odling-Smee, J., Laland, K. N., & Feldman, M. W. (1996). Niche Construction. *The American Naturalist*, *147*(4), 641–648.

Odling-Smee, J., Laland, K. N., & Feldman, M. W. (2003). *Niche Construction: The Neglected Process in Evolution*. Princeton: Princeton University Press.

Osherson, D. N., & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition*, *9*(1), 35–58.

Parsons, T. (1937). *The Structure of Social Action*. *Social Action, New York: Free Press* (Free Press., p. 368). New York: McGraw-Hill.

Penke, L., Denissen, J. J. A., & Miller, G. F. (2007). The Evolutionary Genetics of Personality. *European Journal of Personality*, *21*(5), 549–587. doi:10.1002/per.629

Philippi, T., & Seger, J. (1989). Hedging one's evolutionary bets, revisited. *Trends in Ecology & Evolution*.

Piaget, J. (1932). *The Moral Judgment of the Child.* (M. Gabain, Trans.) (Free Press.). New York: Simon & Schuster.

Pinker, S. (2011). *The Better Angels of our Nature*. New York: Viking.

Plato. (2003). *The Republic*. London: Penguin Books.

Plomin, R., Owen, M. J., & McGuffin, P. (1994). The genetic basis of complex human behaviors. *Science*, *264*(5166), 1733–1739.

Potts, R. (1996). Humanity's Descent: The Consequences of Ecological Instability. New York: William Morrow & Co.

Potts, R. (1998). Variability selection in hominid evolution. *Evolutionary Anthropology: Issues, News, and Reviews*, *7*(3), 81–96. doi:10.1002/(SICI)1520-6505(1998)7:3<81::AID-EVAN3>3.0.CO;2-A

Preston-Mafham, K., & Preston-Mafham, R. (1996). *The Natural History of Spiders*. Ramsbury: Crowood Press.

Prinz, J. (2007). *The Emotional Construction of Morals*. Oxford: Oxford University Press.

Pusey, A. E., & Packer, C. (1997). The Ecology of Relationships. In J. R. Krebs & N. B. Davies (Eds.), *Behavioural Ecology: An Evolutionary Approach* (4th Editio.). New Jersey: Wiley-Blackwell.

Rainey, P. B., & Travisano, M. (1998). Adaptive radiation in a heterogeneous environment. *Nature*, *394*, 69–72.

Rawls, J. (1972). *A Theory of Justice* (Paperback.). Oxford: Oxford University Press.

Richerson, P. J., & Boyd, R. (2001). The evolution of subjective commitment to groups: A tribal instincts hypothesis. In R. M. Nesse (Ed.), *The Evolution of Subjective Commitment*.

Richerson, P. J., & Boyd, R. (2005). *Not By Genes Alone*. Chicago: University of Chicago Press.

Richerson, P. J., Boyd, R., & Henrich, J. (2002). Cultural Evolution of Human Cooperation. In *Genetic and Cultural Evolution of Cooperation* (pp. 1–33).

Ridley, M. (1996). *The Origins of Virtue*. London: Penguin Books.

Ripke, S., Wray, N. R., Lewis, C. M., Hamilton, S. P., Weissman, M. M., Breen, G., … Sullivan, P. F. (2012). A mega-analysis of genome-wide association studies for major depressive disorder. *Molecular Psychiatry*, *18*(4), 497–511. doi:10.1038/mp.2012.21

Rogers, A. R. (1988). Does biology constrain culture? *American Anthropologist*.

Rokeach, M. (1968). *Beliefs, Attitudes and Values: A Theory of Organization and Change* (p. 214). San Francisco: Jossey-Bass.

Rousseau, J.-J. (1755). *A Discourse on the Origin and the Foundation of Inequality Among Mankind* (Cosimo Cla.). New York: Cosimo.

Ruse, M., & Wilson, E. O. (1986). Moral Philosophy as Applied Science. *Philosophy*, *61*(236), 173. doi:10.1017/S0031819100021057

Sagan, C. (1980). *Cosmos*. New York: Random House.

Sartre, J.-P. (1956). *Being and Nothingness*. New York: Washington Square Press.

Sayre, N. F. (2008). The Genesis, History, and Limits of Carrying Capacity. *Annals of the Association of American Geographers*, *98*(1), 120–134. doi:10.1080/00045600701734356

Schoener, T. W. (1971). Theory of feeding strategies. *Annual Review of Ecology and Systematics*.

Scott, J. (2000). Rational Choice Theory. In *Understanding Contemporary Society: Theories of The Present* (Vol. 50).

Selten, R. (1975). Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*.

Shafer-Landau, R. (1994). Ethical disagreement, ethical objectivism and moral indeterminacy. *Philosophy and Phenomenological Research*, *54*(2), 331–344.

Shafer-Landau, R. (2003). *Moral Realism: A Defense*. Oxford: Clarendon Press.

Shweder, R. A., & Haidt, J. (1993). Commentary to Feature Review: The Future of Moral Psychology: Truth, Intuition, and the Pluralist Way. *Psychological Science*, *4*(6), 360–365.

Shweder, R. A., Much, N. C., Mahapatra, M., & Park, L. (1997). The "Big Three" of Morality (Autonomy, Community, Divinity) and the "Big Three" Explanations of Suffering. In A. Brandt & P. Rozin (Eds.), *Morality and Health*. New York: Routledge.

Sidanius, J., & Kurzban, R. (2003). Evolutionary approaches to political psychology. *Oxford Handbook of Political Psychology*.

Sinnott-Armstrong, W. (2006). *Moral Skepticisms*. Oxford: Oxford University Press.

Sinnott-Armstrong, W. (Ed.). (2008). *Moral Psychology, Volume 1: The Evolution of Morality: Adaptations and Innateness*. Cambridge: The MIT Press.

Sinnott-Armstrong, W. (2009). Mixed-Up Metaethics. *Philosophical Issues*, *19*(1), 235–256.

Skyrms, B. (2001). The Stag Hunt. *Proceedings and Addresses of the American Philosophical Association*, *75*(2), 31–41.

Skyrms, B. (2004). *The Stag Hunt and the Evolution of Social Structure*. *Economics and Philosophy* (Vol. 22). Cambridge: Cambridge University Press. doi:10.1017/S026626710621112X

Slade, R. W., & McCallum, H. I. (1992). Overdominant vs. Frequency-Dependent Selection at MHC Loci. *Genetics*, *132*, 861–862. doi:10.2460/javma.238.4.424

Smith, E. A. (2011). Endless forms: human behavioural diversity and evolved universals. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *366*(1563), 325–32. doi:10.1098/rstb.2010.0233

Smith, K. (2002). *Genetic Polymorphism and SNPs. Genomics and Proteomics* (pp. 1–13).

Smith, M. (1994). *The Moral Problem*. Oxford: Wiley-Blackwell.

Spencer, H. (1883). *The Principles of Sociology*. New York: D. Appleton and Company.

Sripada, C. S. (2005). Punishment and the strategic structure of moral systems. *Biology and Philosophy*, 767–789. doi:10.1007/s10539-004-5155-2

Sripada, C. S. (2007). Adaptationism, Culture and the Malleability of Human Nature. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The Innate Mind: Foundations and the Future*. Oxford: Oxford University Press.

Sripada, C. S., & Stich, S. (2004). Evolution, Culture and The Irrationality of the Emotions. In D. Evans & P. Cruse (Eds.), *Emotion, Evolution and Rationality*. Oxford: Oxford University Press.

Sripada, C. S., & Stich, S. (2005). A Framework for the Psychology of Norms. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The Innate Mind: Culture and Cognition* (Vol. II, pp. 1–40). Oxford: Oxford University Press.

Stearns, S. (1989). The evolutionary significance of phenotypic plasticity. *Bioscience*.

Sterelny, K. (2003). *Thought in a Hostile World*. Oxford: Blackwell Publishing.

Sterelny, K. (2007). Social Intelligence, Human Intelligence and Niche Construction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 719–30. doi:10.1098/rstb.2006.2006

Sterelny, K. (2010). *The Evolved Apprentice Essay*. *On the Human*. Retrieved from http://onthehuman.org/2010/08/the-evolved-apprentice/

Sterelny, K. (2012). *The Evolved Apprentice: How Evolution Made Humans Unique*. Cambridge: MIT Press.

Stevenson, C. L. (1937). The Emotive Meaning Of Ethical Terms. *Mind*, *46*(181), 14.

Stevenson, C. L. (1950). The Emotive Conception of Ethics and its Cognitive Implications. *The Philosophical Review*, *59*(3), 291. doi:10.2307/2181986

Storm, I., & Wilson, D. S. (2009). Liberal and Conservative Protestant Denominations as Different Socioecological Strategies. *Human Nature*, *20*(1), 1–24. doi:10.1007/s12110-008-9055-z

Street, S. (2006). A Darwinian Dilemma for Realist Theories of Value. *Philosophical Studies*, *127*(1), 109–166. doi:10.1007/s11098-005-1726-6

Tavits, M. (2005). Causes of Corruption: Testing Competing Hypotheses. In *Joint Sessions of Workshops* (pp. 1–32).

Tersman, F. (2006). *Moral disagreement*. Cambridge: Cambridge University Press.

Tetlock, P. E. (1983). Cognitive style and political ideology. *Journal of Personality and Social Psychology*, *45*(1), 118–126. doi:10.1037/0022-3514.45.1.118

Tetlock, P. E. (2002). Social Functionalist Frameworks for Judgment and Choice: Intuitive Politicians, Theologians, and Prosecutors. *Psychological Review*, *109*(3), 451– 471. doi:10.1037//0033-295X.109.3.451

Tinbergen, N., Broekhuysen, G. J., Feekes, F., Houghton, J. C. W., Kruuk, H., & Szulc, E. (1962). Egg shell removal by the black-headed gull, Larus ridibundus L.; a behaviour component of camouflage. *Behaviour*, 74–117.

Tooby, J., & Cosmides, L. (1990). The Past Explains the Present. *Ethology and Sociobiology*, *11*(4-5), 375–424. doi:10.1016/0162-3095(90)90017-Z

Tooby, J., & Cosmides, L. (2005). Conceptual foundations of evolutionary psychology. In D. M. Buss (Ed.), *The handbook of evolutionary psychology* (pp. 5–67).

Tooby, J., & Cosmides, L. (2009). Conceptual foundations of evolutionary psychology. *Philosophy of Biology: An Anthology*, 375.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, *46*(1), 35–57.

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

Turkheimer, E. (2000). Three laws of behavior genetics and what they mean. *Current Directions in Psychological Science*, *9*(5), 160–164.

Turner, S. P. (1995). Durkheim's "The Rules of Sociological Method": Is It a Classic? *Sociological Perspectives*, *38*(1), 1–13.

Van Dover, C. L., German, C. R., Speer, K. G., Parson, L. M., & Vrijenhoek, R. C. (2002). Evolution and biogeography of deep-sea vent and seep invertebrates. *Science (New York, N.Y.)*, *295*(5558), 1253–7. doi:10.1126/science.1067361

Van Valen, L. (1973). A New Evolutionary Law. *Evolutionary Theory*, *1*(1), 1–30.

Verweij, K. J. H., Vinkhuyzen, A. A. E., Benyamin, B., Lynskey, M. T., Quaye, L., Agrawal, A. A., … Medland, S. E. (2013). The genetic aetiology of cannabis use initiation: a meta-analysis of genome-wide association studies and a SNP-based heritability estimation. *Addiction Biology*, *18*(5), 846–50. doi:10.1111/j.1369-1600.2012.00478.x

Von Luck, H. (1989). *Panzer Commander*. London: Cassell.

Walter, A. (2006). The Anti-naturalistic Fallacy: Evolutionary Moral Psychology and the Insistence of Brute Facts. *Evolutionary Psychology*, (4), 33–48.

Webb, B. T., Guo, A.-Y., Maher, B. S., Zhao, Z., van den Oord, E. J., Kendler, K. S., … Hettema, J. M. (2012). Meta-analyses of genome-wide linkage scans of anxiety-related phenotypes. *European Journal of Human Genetics*, *20*(10), 1078–84. doi:10.1038/ejhg.2012.47

West, S. A., Griffin, A. S., & Gardiner, A. (2007). Social semantics: how useful has group selection been? *Journal of Evolutionary Biology*, *21*(November 2007), 374–385. doi:10.1111/j.1420-9101.2007.01458.x

West, S. A., Griffin, A. S., & Gardner, A. (2007). Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology*, *20*(2), 415–32. doi:10.1111/j.1420-9101.2006.01258.x

Westermarck, E. (1906). *The origin and development of the moral ideas (Vols. 1 and 2). London* (Vol. 8). Freeport, New York: Books for Libraries Press.

Westermarck, E. (1932). *Ethical Relativity*. Westport: Greenwood Press.

Whelan, R. J. (1995). *The Ecology of Fire*. Cambridge: Cambridge University Press.

Whitehead, A. N. (1919). *The Concept of Nature*. New York: Cosimo.

Wilson, D. S. (1994). Adaptive genetic variation and human evolutionary psychology. *Ethology and Sociobiology*, *15*(4), 219–235. doi:10.1016/0162-3095(94)90015-9

Wilson, D. S. (2002). *Darwin's Cathedral*. Chicago: University of Chicago Press.

Wilson, D. S. (2005). Testing Major Evolutionary Hypotheses about Religion with a Random Sample. *Human Nature*, *16*(4), 382–409.

Wilson, D. S., Clark, A. B., Coleman, K., & Dearstyne, T. (1994). Shyness and boldness in humans and other animals. *Trends in Ecology & Evolution*, *9*(11), 442–6. doi:10.1016/0169-5347(94)90134-1

Wilson, D. S., & Sober, E. (1994). Reintroducing group selection to the human behavioral sciences. *Behavioral and Brain Sciences*, *17*(04), 585–654. doi:10.1017/S0140525X00036104

Wilson, E. O. (1975). *Sociobiology: The New Synthesis*. Cambridge: Harvard University Press.

Wilson, E. O. (1980). The Ethical Implications of Human Sociobiology. *The Hastings Center Report*, *10*(6), 27–29.

Wolf, M., & Weissing, F. J. (2010). An explanatory framework for adaptive personality differences. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *365*(1560), 3959–68. doi:10.1098/rstb.2010.0215

Wong, D. (1984). *Moral Relativity*. Berkeley: University of California Press.

Wong, D. (1991). Relativism. In P. Singer (Ed.), *A Companion to Ethics*. Oxford: Blackwell Publishing.

Wong, D. B. (2006). *Natural Moralities: A Defense of Pluralistic Relativism*. Oxford: Oxford University Press.

Wrangham, R. (2009). *Catching Fire: How Cooking Made Us Human*. London: Profile Books.

Wright, L. (1973). Functions. *Philosophical 1Review*, *82*(2), 139–168.

Wu, J., & Axelrod, R. (1995). How to cope with noise in the iterated prisoner's dilemma. *Journal of Conflict Resolution*.

Zahavi, A. (1975). Mate Selection - A Selection for a Handicap. *Journal of Theoretical Biology*, *53*(205-214).

Zahavi, A., & Zahavi, A. (1997). *The Handicap Principle: A Missing Part of Darwin's Puzzle*. New York: Oxford University Press.

Zaller, J. R. (1992). *The Nature and Origins of Mass Opinion*. Cambridge: Cambridge University Press.