EVOLVED VALUE AND THE FOUNDATIONS OF ETHICAL THEORY

by

GEORGE M. FELIS

(Under the Direction of Scott A. Kleiner)

ABSTRACT

Ethical theories seem always to be rooted in some underlying theory of human nature, explicit or implicit: Understanding what kind of creatures humans are would appear to be a necessary precondition for drawing conclusions about how humans ought to live. In the 150 years since the publication of Darwin's *Origin of Species*, humanity's collective understanding of its own nature has grown increasingly rooted in evolutionary biology. However, a scientific theory of human nature is widely presumed to provide a basis only for factual claims about human nature, from which no value conclusions can be drawn: This presumption, if true, would permit only descriptions of evolved ethical behaviors and would deny the possibility of producing a genuinely prescriptive ethical theory rooted in evolutionary biology.

While I acknowledge the difficulty of any argument which attempts to bridge the gap between facts and values, I believe that this difficulty can be overcome. In my dissertation, I develop the foundation for a prescriptive evolutionary ethical theory as follows: The arguments offered by Aristotle, Kant and Mill for the foundations of virtue ethics, deontology and utilitarianism respectively all attempt to bridge the fact-value gap in the same basic fashion – by identifying what is, as a matter of fact, of intrinsic value to each and every human being. These fact-value bridging claims serve as foundational normative premises from which prescriptive conclusions can be justifiably derived, culminating in the universal prescriptive claims of their respective complete ethical theories. From an evolutionary perspective, the fitness benefits (and costs) to a given organism of its various possible circumstances and activities effectively comprise what is of intrinsic value (and disvalue) to that organism. Higher-level selective processes such as kin selection, reciprocal altruism and group selection can broaden the initially self-regarding character of what is of value to a given organism in a way that includes selected other organisms. In humans, the mechanisms of cultural selection expand what is of value to an individual human to include every other human, at least in a certain respect – which constitutes the core normative premise foundation for a prescriptive evolutionary ethical theory.

INDEX WORDS: Ethics, Ethical theory, Evolution, Evolutionary ethics,
Fact value problem, Is ought problem, Metaethics, Teleology

EVOLVED VALUE AND THE FOUNDATIONS OF ETHICAL THEORY

by

GEORGE M. FELIS

B.A., Miami University, 1989

M.A., Miami University, 1997

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2009

EVOLVED VALUE AND THE FOUNDATIONS OF ETHICAL THEORY

by

GEORGE M. FELIS

Major Professor:	Scott A. Kleiner
Committee:	O. Bradley Bassler Robert G. Burton Victoria Davion

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2009

ACKNOWLEDGEMENTS

I would like to thank the following University of Georgia Department of Philosophy faculty for their support: Professor Emeritus Scott A. Kleiner for his willingness to serve as my major professor, even after his retirement; Professor Emeritus Robert G. Burton for years of recommendation letters and general positive reinforcement as Department Head; Professor Vicky Davion for serving on my committee despite her time-consuming duties as the new Department Head; and Professor Brad Bassler for more fascinating classes and wonderful conversations than I can easily recall. I would also like to thank the entire Franklin College Office of Academic Advising, especially Advising Coordinator Laura Dowd; the supportive, flexible, and always-friendly work environment I have enjoyed as an academic advisor have been instrumental to my completion of this project.

TABLE OF CONTENTS

PREFACE

"Philosophy, like all other studies, aims primarily at knowledge. The knowledge it aims at is the kind of knowledge which gives unity and system to the body of the sciences, and the kind which results from a critical examination of the grounds of our convictions, prejudices, and beliefs. But it cannot be maintained that philosophy has had any very great measure of success in its attempts to provide definite answers to its questions. If you ask a mathematician, a mineralogist, a historian, or any other man of learning, what definite body of truths has been ascertained by his science, his answer will last as long as you are willing to listen. But if you put the same question to a philosopher, he will, if he is candid, have to confess that his study has not achieved positive results such as have been achieved by other sciences. It is true that this is partly accounted for by the fact that, as soon as definite knowledge concerning any subject becomes possible, this subject ceases to be called philosophy, and becomes a separate science. The whole study of the heavens, which now belongs to astronomy, was once included in philosophy; Newton's great work was called 'the mathematical principles of natural philosophy'. Similarly, the study of the human mind, which was a part of philosophy, has now been separated from philosophy and has become the science of psychology. Thus, to a great extent, the uncertainty of philosophy is more apparent than real: those questions which are already capable of definite answers are placed in the sciences, while those only to which, at present, no definite answer can be given, remain to form the residue which is called philosophy."

– Bertrand Russell, *Problems of Philosophy*, Ch. XV [p.239-240]

CHAPTER 1

INTRODUCTION

Section 1: Motivations, goals, and challenges for this project

> ...All the same, we must look more closely at the matter, since what is at stake is far from insignificant: it is how one should live one's life.
> – Plato, *The Republic*, (Book I) [Plato, p.40]

How should one live one's life? That was the question that motivated Socrates' inquiry in Plato's *Republic*, and two dozen centuries or so later the same question still motivates extensive philosophical inquiry – not least because there is little general agreement on any of the answers offered to date. Indeed, there is little agreement on how to go about answering the question, on whether and how we could discern if a given answer is true, or on whether the question can be truly answered at all. Such is the unsettled state of ethical theory.

In *The Problems of Philosophy* (quoted in my Preface), Bertrand Russell contended that such unsettled principles of inquiry are not only inevitable in philosophy, they are diagnostic of the field: When basic questions of methodology and evidence are resolved and foundations for further progress have been widely accepted, an area of inquiry passes from the realm of philosophy and becomes a new science. Philosophy, on Russell's view, is in some sense the residue of inquiry, the discipline of unanswered questions and unresolved disputes. Philosophy consists in those subjects where even the matter of whether the right questions are being asked often remains unresolved, let alone matters such as the kind of evidence needed to answer questions and the methods by which such evidence is to be gathered and evaluated. [Russell, p.239-240]

Russell's perspective on the unsettled and open-ended character of philosophical inquiry is useful in two ways: It offers hope that this state of affairs need not be permanent, and it suggests a way forward. We can look to those areas of inquiry where there are settled methodologies and well-established bodies of knowledge, and use those to inform and constrain philosophical inquiry wherever we can. We can expose the flaws in bad questions and frame better questions, find more fruitful lines of inquiry and abandon dead-ends. We can even, if we are very fortunate, establish firmer foundations upon which we can move substantially forward.

Russell's perspective does risk a sort of reflexive absurdity, since the nature and bounds of philosophy would seem to be among those unsettled areas of ongoing dispute. Nevertheless, the accelerating frequency with which philosophers turn to science to inform and constrain philosophical investigation suggests that many philosophers embrace something like Russell's view, at least tacitly: Almost as soon as there were computers, philosophers of mind turned to computer scientists in hopes that "electronic brains" might offer insights into the workings of the human mind. More recently, philosophers of mind have paid almost as much attention to neurophysiology and functional magnetic resonance imaging as argument and analysis in their quest to understand and solve the "hard problem" of consciousness. For an example outside the philosophy of mind, one can hardly read modern writings on the metaphysics of time without reading as many references to Einstein and space-time manifolds as to more traditional philosophical analyses of time. And so on.

But my aim here is not to discuss historical and current intersections of science and philosophy, nor to defend Russell's particular perspective on the relationship between the two. Rather, I raise these background issues to convey some of the context of and motivations for my chosen approach to the broad question of how one ought to live: The overall goal of this dissertation is to determine whether and how science can be used to inform and constrain ethical theory. More specifically, I intend to develop an understanding of human nature grounded in evolutionary biology which establishes firmer foundations for ethical theory than any of the traditional contenders. Throughout this dissertation, I will take it as a given that evolutionary biology (writ large) is an accurate explanation and true description of the operations of the living world. Those who have doubts on that matter, I can only refer to a work by someone more qualified than I – perhaps Jerry Coyne's *Why Evolution Is True* [Coyne, 2009]. However, the overwhelming empirical evidence for evolution can only help establish firmer foundations for ethical theory if used very carefully – for reasons that will become clear below.

Human nature is a natural starting point for this project, both because it is an area where science has radically changed our views in a relatively short time and because it is intimately tied to ethical theory: To make any argument about how humans ought to live, it would seem necessary to have some understanding of what sort of creatures humans are. It is difficult to imagine how any prescriptive ethical (or political) theory could be intelligible if it were not rooted in some conception of human nature. Even the insistence that there is no such thing as human nature – a claim with some current popularity, although in many different and often mutually exclusive variations – is in fact a claim about human nature, and one that has definite consequences for how its proponents understand and generate ethical and political prescriptions.

Inquiry into human nature is itself a source of considerable dispute and many open questions, of course. But insofar as accounts of human characteristics and capacities are amenable to scientific investigation, the disputes are potentially resolvable, and the open questions answerable. I do not intend to imply that the ongoing disputes and open questions of ethical theory proper are inherently or necessarily irresolvable, or that none have ever been settled – but the science of human nature certainly has a much better track record in this regard than moral philosophy, and in a handful of generations rather than a few dozen centuries.

In the one hundred and fifty years since the publication of Charles Darwin's *On the Origin of Species*, understanding human nature scientifically has in large measure meant understanding human nature as a product of evolution by natural selection. This approach recognizes that human characteristics and capacities form a continuum with those of other organisms, and has yielded many insights into the biological bases of behaviors we typically think of as ethical, whether observed in ourselves or in other organisms. For example, the ability to understand fairness – or at least to recognize and react negatively to manifestly unfair treatment of oneself – would seem to be shared by many other social animals: not just our closest cousins the chimpanzees [de Waal, 1991], but much more distantly related monkeys [Brosnan & de Waal, 2003], and even domesticated canines [Range *et al*, 2009]. The descriptive and explanatory successes of evolutionary accounts of ethical behavior can be misleading, however, because they tell us *how organisms do behave*, not *how they ought to behave*.

Broadly speaking, 'ought' claims fall into two categories of value claims: 'normative claims' about how the world ought to be, i.e. claims about what is good or valuable; and 'prescriptive claims' about how people in general ought to behave, or about what some particular person ought to do in some circumstance. Sometimes this distinction is characterized as 'the good' (normative claims) and 'the right' (prescriptive claims). In this vocabulary, the right follows from the good. Naturally enough, what one concludes the right action to be depends on what one considers to be good/valuable.

It is crucial to remember that judgment concerning what constitutes ethical behavior in the first place – fairness, reciprocity, altruism, and so on – is already informed by normative convictions. To return to my chosen example, demonstrating that humans and other social animals have an instinct for fairness does not of itself tell us that fairness is good. Rather, psychologists, primatologists and ethologists designed the experiments cited above to evaluate whether and to what extent other animals would recognize and react to unfair treatment precisely because we already value fairness: Humans generally see fairness as having great practical and ethical importance in human society. Moreover, because the capacity to recognize unfairness and the inclination to oppose it seem to be products of natural selection (and not just in humans), it seems likely that the inclination to believe fairness is good is itself the product of natural selection – which, again, does nothing to establish that fairness *is* good.

More generally, value judgments determine what we designate as ethical or unethical behaviors well in advance of any descriptions and explanations for such behaviors which might result from scientific investigation. In my chosen example, the judgment that fairness is valuable and therefore worthy of study constitutes an assumed normative claim: Thus, any evidence we find that humans generally value fairness (or disvalue unfairness) cannot be evidence that fairness is valuable, because that is the assumption we started from. To draw a normative conclusion from such research would be circular, and to use such an assumed normative claim as a basis for further prescriptive conclusions would only compound the error.

Taken together, these considerations pose something of a dilemma. On the one hand, the need to understand what kind of creatures humans are before making any claims about how humans ought to behave seems very clear. On the other hand, it is not at all clear how prescriptive conclusions about how humans ought to behave can be reached based upon descriptive or explanatory claims regarding how humans actually do behave – especially when decisions about what requires special description and explanation as "ethical" behavior already depend upon prior normative claims. Explanations for human behavior grounded firmly in scientific inquiry – which ultimately means explanations for human behavior grounded in evolutionary biology – may be very useful in producing a factual, purely descriptive account of ethical behavior and capacities. However, determining which human behaviors even count as ethical or unethical – let alone giving an account of *why* they are ethical or unethical, as opposed to why they evolved in our ancestors – seems well beyond the scope of the scientific inquiry.

Since prescriptive claims are, generally speaking, based on normative claims, discussions of this problem usually focus on the gap between descriptive claims and normative claims. Because it relies on empirical evidence from the world as it is, science can only describe the world as it is, which can also include causal narratives which tell how the world came to be as it is (explanatory claims). Based on such a causal narrative, science can even tell us that some state of affairs might have been otherwise, but for this or that causal factor. But science cannot tell us thereby that it *ought* to have been otherwise, or that it *should* be otherwise, or that it would be *better* (or *worse*) if it were otherwise. The latter are normative claims – claims which require some standards or norms to which the world as it is can be compared. It would seem impossible that those standards could be found in the world as it is, or such comparisons would only and always conclude that the world is exactly as it ought to be. Such an assumption is so clearly flawed – not to mention being at odds with the very meaning of 'ought' and related concepts – that it has been labeled a fallacy, usually called the "is-ought fallacy."[1]

The distinction between claims about the world as it is and claims about how the world ought to be is also sometimes made in terms of 'factual claims' versus 'value claims.' Factual claim premises do not support value claim conclusions, a problem that is often called the 'fact-

[1] It has also been mistakenly referred to as the "naturalistic fallacy," but G.E. Moore coined that phrase in reference to a different (although closely related) issue, which I will discuss in Chapter 5.

value gap' or 'fact-value problem.' While this fact-value problem is not at its heart any different from the gap between descriptive and normative claims that leads to the is-ought fallacy, the alternate wording emphasizes another aspect of the problem: One conception of 'facts' is that they are truth-makers for factual claims. That is, a proposed factual claim is true if and only if it refers to some state of affairs that obtains in the world. Even without getting into an exhaustive discussion of the ontological, metaphysical and semantic issues at stake here – what a 'state of affairs' actually is, what it means for a claim to 'refer,' and what 'truth-making' amounts to – the fact-value distinction raises compelling questions about the nature of value, questions usually gathered together and addressed under the label *metaethics*: What is it that value claims refer to that makes them true or false? Are there moral facts? If there are moral facts, they must differ from other sorts of facts, since they cannot simply be obtaining states of affairs. How are they different, and why? And if there are no moral facts, what are truth-makers for value claims? Does it even make sense to say that a value claim is true or false? If not, does a value claim express any sort of belief, or is it some other kind of utterance?

Moreover, value claims are multifaceted. The emphasis throughout this project is on moral value claims – so much so that I will often drop the word "moral" – but not all normative or prescriptive claims are moral value claims (claims about the good or the right). For example, one can discuss the norms of and prescriptions for good reasoning. Similarly, aesthetic claims are also classified as value claims – and it may be that all of the questions that conclude the previous paragraph have one set of answers for moral value claims, a different set of answers for non-moral value claims, and a still different third set of answers for aesthetic claims. Aside from all these differences, there might also be differences even within such categories: Some moral value claims may be statements of belief which can be judged true or false by some means, and others may be nothing more than emotive expressions or statements of personal preference. That is, it may be that some kinds of moral claims can be assessed as true or false, and others not.

Such fundamental questions about the nature of value must be addressed at some level in the course of this project, insofar as any proposed ethical theory must take positions with regard to many if not all of these issues, and my goal is to develop the foundations for an ethical theory. However, a survey of the competing arguments and fine distinctions that characterize the history of metaethics would not move me even a step forward towards that goal. So, instead of attempting to resolve these metaethical problems in advance, I will raise and address them as they arise in the course of my argument. But if I am to bring scientific findings to bear on ethical theory at all, the problem that clearly must be addressed first is the fact-value problem itself.

Section 2: Hume's fact-value problem

> In every system of morality, which I have hitherto met with, I have always remark'd, that the author proceeds for some time in the ordinary ways of reasoning, and establishes the being of a God, or makes observations concerning human affairs; when of a sudden I am surpriz'd to find, that instead of the usual copulations of propositions, *is*, and *is not*, I meet with no proposition that is not connected with an *ought*, or an *ought not*. This change is imperceptible; but is however, of the last consequence. For as this *ought*, or *ought not*, expresses some new relation or affirmation, 'tis necessary that it shou'd be observ'd and explain'd; and at the same time that a reason should be given; for what seems altogether inconceivable, how this new relation can be a deduction from others, which are entirely different from it.
>
> – David Hume, *Treatise of Human Nature* (Book III) [Hume, p.302]

Hume's skeptical assessment of ethical theory, quoted above, is one of the more influential and controversial passages in the history of moral philosophy. However, it is not my intent to trace this passage's influence, nor to discuss the controversies over the interpretation and application of Hume's insight. What the passage says on a straightforward reading is quite substantial and worth addressing in its own right: Factual propositions, claims about the way the world is, are fundamentally different from value propositions, claims about the way the world ought to be or actions people ought to take – and one cannot move blithely from factual premises to value conclusions. It is worth noting that even the arch-skeptic Hume does not claim that it is *impossible* to base value conclusions on factual premises, or at least he doesn't claim so here. He does flatly state that one cannot possibly move from factual premises to value conclusions by *deduction* – or at least, he calls such a deduction "inconceivable." Because a connection between fact and value cannot be a matter of simple deduction, Hume suggests that it must be *justified* – that it must be "observ'd and explain'd," and "a reason should be given." If one is to draw value conclusions from factual premises, one must make some sort of additional argument that connects the two different sorts of claim.[2]

If one takes Hume's characterization of the fact-value problem seriously, every ethical theory faces an obstacle at a very basic level: The principle or principles which constitute the conclusion of any well-developed ethical theory – Kant's categorical imperative, the principle of utility, etc. – consist in a set of prescriptive value claims; sentences copulated by 'ought' or some equivalent thereof, to use Hume's phrasing. The only evident solution to the problem Hume raises is to reason towards these prescriptive value claims from premises that include at least some prior value claims, which would be either other prescriptive claims or normative claims.[3]

[2] Despite what I take to be Hume's fairly clear articulation of his position, some philosophers have treated Hume's claim that deductions of 'ought' claims from 'is' claims are impossible as the only relevant or interesting issue he raises in this passage. For example, in his famous paper "How to Derive 'Ought' From 'Is'," John Searle treats is-to-ought arguments as a sort of enthymeme, and claims that the problem is solved by supplying the missing fact-value-bridging premise [Searle, 1964]: This approach seems to ignore the major problem of how such a fact-value-bridging premise is itself to be justified, which problem is exactly what I take Hume to be emphasizing when he says that an ought claim "expresses some new relation or affirmation..." The new relation would lie in that fact-value-bridging premise, and making that premise explicit does not constitute giving a reason for it. For a useful discussion of the distinction between justification and deduction in this context, see Ken Witkowski's "The 'Is-Ought' Gap: Deduction or Justification?" [Witkowski, 1975].

[3] For the sake of completeness, I should note that I am taking for granted that the primary obstacle in this context is basing value claim conclusions on factual claim premises, and setting aside for now any

But those value claim premises in turn stand in need of justification, one must suppose, which would require still more prior value claims. Therefore, Hume's fact-value problem has but three possible "solutions," to use the word somewhat loosely: (1) to settle for an infinite regress of value claims, (2) to construct a circular argument where the value conclusion is included amongst the value premises in some way, or (3) to take some value premise or premises to be basic principles, the foundational value claim(s) from which all other value claims spring.

The first two of these – an infinite regress of premises requiring further justification, and an argument which includes the conclusion amongst the premises – constitute justification problems every bit as grave as the unjustified, unexplained move from 'is' to 'ought' Hume criticizes, if not more so. That leaves only one plausible solution: If one is to justify any ethical theory, one must take some value claim or claims to be foundational premises. One cannot simply declare or assert such a foundational value premise(s) and be done with it; one must make an argument in favor of the foundational value premise(s) in question. That is, one must justify one's choice of a particular underlying value claim (or set of value claims). But if one is to avoid the problem of value claims depending upon other value claims which depend on still other value claims *ad infinitum*, one must justify one's foundational value claim(s) on non-value grounds. And what might "non-value grounds" be if not factual claims? It seems as if the only plausible solution to the fact-value problem requires that the fact-value problem already be solved.

One might object at this point that there appears to be another solution: Rather than justifying a foundational value claim or claims on non-value grounds, one could discover foundational value claim(s) which are self-evident or self-justifying in some way. Moral

complexities or problems that may be associated with drawing prescriptive conclusions from normative or prescriptive premises.

intuitionists and empiricists have developed arguments along these lines[4], but such positions are difficult to defend from the charge that they ignore the distinction between appearance and truth: That is, appealing to people's moral intuitions or observed moral sentiments only establishes people's *opinions* about what ought to be done (or their feelings about what ought to be done, or some other subjective cognition), not what *actually* ought to be done. Defense against this criticism would seem to require finding some way to justify the foundational value claim or claims which such theories take to be self-evident or self-justifying – which brings the argument back around to the same justification problem as before.

One might be tempted here to say that this reflects a general epistemological problem rather than any obstacle specific to value claims. What are the self-justifying or self-evident premises that can be used to justify factual claims, to distinguish what *is* the case from what merely *appears* to be the case? Without delving too far into epistemological matters, I will simply point out that there is at least one crucial difference: In disputes over matters of fact, one can in some sense or another (depending on one's epistemological preferences) appeal to the world itself. We can go see for ourselves. We can even conduct randomized, double-blind, reproducible, controlled experiments designed to remove as many natural human biases and potential delusions as possible from the process of seeing for ourselves.

In contrast, the search for justification based on moral intuitions or sentiments runs into serious obstacles upon the first conflict between one person's intuitions or sentiments and another's: In disputes over matters of value, one cannot appeal to the world as one can in disputes over matters of fact. The world as it is differs substantially from the world as we think it ought to

[4] It is worth noting that Hume himself follows this route: The title of the section which follows the paragraph cited above is "Moral distinctions deriv'd from a moral sense." [Hume, p. 302] Whether or not there is a genuine difference between intuitionism and empiricism about moral matters might depend upon how one describes and accounts for the imputed moral sense, but any further discussion along those lines would stray from the matter at hand.

be, or there would be no motivation for ethical inquiry in the first place. Even without dispute between competing claims – that is, even if a sufficiently careful analysis of human moral intuitions or sentiments revealed at least some value claims on which there is perfect universal agreement – mere agreement is surely not the same as genuine justification.

As a further objection, I suppose someone might protest that matters of empirical fact are also based on a sort of agreement, insofar as we can only appeal to our experiences of the world and cannot appeal to the world in itself. But our collected and compared experiences of the world as it is can (and regularly do) conflict with our beliefs about the world prior to collecting and comparing experiences, and such conflicts motivate and justify changing our beliefs. We cannot collect and compare *experiences* of the world as it ought to be, we can only compare our *beliefs* about how we think the world ought to be. Thus, we have no independent basis for comparison, no possible experiences which we could collect and compare to our prior beliefs, and thus no possible conflicts with new experiences to motivate and justify changing our beliefs. So even if one is dubious about the extent to which the empirical foundations for factual claims are self-evident or self-justifying – even if one is the most radical epistemological skeptic imaginable – one can see that value claims are further still from any such self-evidence, and Hume's problem remains.

Despite the apparent inescapability of this impasse, ethical theory does not appear to have been entirely undone by Hume's formulation of the fact-value problem. Philosophers before and since Hume have offered extensive justifications for ultimate prescriptive principles and the normative claims on which they are based, and have taken themselves to have done so successfully. Whether those justifications are adequate is another matter, but many thinkers besides those who offered the arguments have taken them to be convincing, or at least plausible.

So for the sake of argument, I will take the minimally charitable position that ethical theorists are not all deluded, that they have not each and every one made the unjustified and mysterious deduction of values from facts that Hume rightly criticizes. If not, then the next question to ask is whether and how moral philosophers have attempted to justify at least some value claim(s) on non-value grounds.

I am not assuming that all moral philosophers since Hume – and still less those before Hume[5] – have explicitly addressed the fact-value problem in these terms, let alone resolved it. But if one takes the fact-value problem to be a serious obstacle facing any ethical theory, one naturally might want to investigate whether and how various respected moral philosophers have – or have not – solved it. That is, one would want to see whether and how the connection between fact and value is made explicit, clear, and plausible in various influential ethical theories. Or to translate that question back into Hume's own words: Since the connection between factual claims and value claims is a "new relation or affirmation, 'tis necessary that it shou'd be observ'd and explain'd; and at the same time that a reason should be given."

In Chapter 2, I develop a general approach to solving Hume's fact-value problem. The solution requires establishing a fact-value bridging claim which meets specific criteria, which claim can serve as the foundation for further claims, and thus as the foundation for an ethical theory. Then I will argue that the three most influential ethical theories – utilitarianism, deontology and virtue theory – each bridge the gap between facts and values with their own versions of this general solution. In Chapter 3, I will follow the model offered by Aristotle's biology-friendly ethical theory to develop a version of this solution firmly rooted in evolution by

[5] The strong division between descriptive and normative claims (and the perception of a wide gap between the two) is a problem of relatively recent philosophical vintage. As I shall discuss later, Aristotle's view of human nature – his view of nature as a whole, for that matter – has no room for this dichotomy. I will also show how Aristotle's view, while it does not really acknowledge the fact-value binary, does offer a solution to Hume's problem if one examines it carefully.

natural selection, thereby developing a naturalized account of value for all organisms. In Chapter 4, I will argue that this account of naturalized value as it is embodied in humans specifically (rather than organisms generally) meets all the criteria established in Chapter 2 for a fact-value bridging claim that can serve as the foundation for an ethical theory, although I will have little to say about what sort of ethical theory might be generated from such a foundation. In Chapter 5, I will conclude by considering a final broad objection to my argument, specifically G.E. Moore's Open Question Argument, and by examining a few of the implications of my conclusion.

CHAPTER 2

BRIDGING THE FACT-VALUE GAP

Section 1: Sketching a solution for the fact-value problem

Prescriptive ethical theories seem always to be rooted in some underlying theory of human nature, explicit or implicit: Understanding what kind of creatures humans are would appear to be a necessary precondition for drawing any conclusions about how humans ought to live.[6] However, to look at any account of human nature as being purely descriptive, as simply a collection of value-neutral factual claims, is surely somewhat naive. Skepticism about the move from 'is' to 'ought' would seem at least partly inspired by the way values are often smuggled into a theory of human nature disguised as facts, thereby avoiding careful scrutiny. Even more commonly, values are not so much disguised as facts, but simply assumed from facts – instances of the is-ought fallacy, defined and discussed in my introduction (see p.7).

However, if addressed explicitly and carefully instead of being smuggled in, and if care is taken to avoid the is-ought fallacy, the notion that some claims about human nature are value-neutral and others are value-laden opens up an intriguing possibility for bridging the fact-value gap: Is it possible for there to be factual claims that are nevertheless value-laden in some sense? In everyday usage, it certainly seems plausible that there are facts about persons which have

[6] I use the cautious phrasing "seem" and "appear" here only to avoid digression. I am fairly certain that it is indeed both an historically universal and logically necessary feature of prescriptive ethical or political theory to be based upon a prior descriptive account of human nature – or, more broadly, a descriptive account of moral agency that includes humans as such agents – but even if I have missed several counter-examples, nothing in the rest of my argument hinges on this universality claim.

prescriptive consequences. For example, it is a fact about me that I value education, which surely implies that I ought not drop out of graduate school to make more time for unicycling.

One might object that some equivocation in the use of 'value' is at work here: My personal preferences or interests, however strong, do not constitute or generate moral value claims of the sort that matter for ethical theory; they do not state or imply any morally binding 'ought' or 'ought not.' Such a dismissal seems too facile, however. Whether and to what extent a being has the capacity to experience pain is a matter of fact, but people in general and moral philosophers in particular (especially utilitarians) tend to think that facts about pain and suffering have clear value implications of great importance. Yet, the value I place on avoiding pain – and the value anyone else places on it – would seem to be just as much a matter of personal interest and preference as the value I place on education, albeit more widely shared. In my case, the value of education has even trumped the value of avoiding pain to some degree, insofar as I have been willing to tolerate considerable pain of various kinds to further my education. So whether particular to individuals or universal to humans, it seems that there can be facts about what is of value (or disvalue) to us that plausibly offer a basis for drawing prescriptive conclusions, and perhaps even moral prescriptive conclusions.

Admittedly, not just any facts about what is of value or disvalue to just anyone will do for the purposes of ethical theory. The ultimate prescriptive principles of ethical theories are usually taken to be universal: Indeed, an 'ought' statement would not accurately be called a 'principle' if it were not generally applicable. So it seems reasonable that the basis for deriving ethical principles that apply to everyone must be facts which are true of everyone – specifically, facts about what is of value (or disvalue) to everyone.

Even presented only as a very loose sketch so far, this line of thinking – one hesitates to call it an argument yet – is not as straightforward as it may seem. In the previous paragraph, I used the phrase "for the purposes of ethical theory" without filling in what those purposes might be. In one sense, there is nothing controversial about that. The purpose of ethical theory is to find answers to the question I posed at the beginning of Chapter 1: How should one live one's life? But any attempt to answer that question is likely to raise the sort of metaethical problems that I mentioned in Chapter 1. In the remaining sections of this chapter, some of those exact questions will be raised as I develop a more detailed argument that facts about what is of value (or disvalue) to us can plausibly serve to resolve the fact-value problem, and look at how such arguments have been used by various ethical theorists.

Section 2: Developing a solution for the fact-value problem

It is a matter of fact whether and to what extent any individual human does value particular states of affairs. For the sake of illustration, I value education; Alice values independence; Bob values financial security. More precisely, we each value states of affairs in which we obtain or advance our education, independence, or financial security. Further, each of us values the states of affairs that the other two value, although not necessarily to the same degree. Those facts about what each of us values provide support for conclusions about what each of us ought to do in order to bring about the states of affairs we value – at least where 'ought' is grounded in the pragmatic norms of instrumental reasoning. For example, graduate study in philosophy is a better choice for me than for Bob, given that I place a higher value on education than he does, and he places a higher value on financial security than I do.

Such claims, which I will refer to as 'bridging claims' hence, bridge the fact-value gap by virtue of their content: Such a claim states a matter of fact, insofar as a sentence like "Alice values independence" is a claim about the world that is either true or false, and there are reasonable means by which its truth or falsity can be determined. (For example, "Go ask Alice!") However, even though they have the *form* of factual claims and even some factual content, such bridging claims also refer to or imply the value *content* of normative claims (if not always moral normative claims), and thus have the potential to support prescriptive claim conclusions.

Of course, the relationship between what is valuable to someone and what they ought to do is not simple. The fact-value bridging normative claims I have been considering here can be generalized to encompass any claim of the form, "state of affairs *X* is of value to *A*." A prescriptive claim with the same referential content would have the form, "*A* ought to take action to bring about state of affairs *X*." Less formally, some state of affairs being of value provides a basis for concluding, other things being equal, that some action or actions ought to be taken (usually by the valuer) in order to bring about that state of affairs. But other things are rarely equal, so the connection between a normative claim premise and a prescriptive conclusion is a good deal more complicated than such a loose description captures, even for pragmatic norms.

The connection between a given normative claim and its related prescriptive claim might appear to be nearly analytic – if a given normative claim is true, the parallel prescriptive claim would seem to follow more or less automatically. However, as the discussion of Bob and me above illustrates, relations among different values can lead to different prescriptive conclusions: Education is valuable to both Bob and me, but Bob values financial security more than education and I value education more than financial security, which leads to different prescriptive conclusions for each of us. Drawing prescriptive conclusions from normative claim premises requires a complex reasoning process which balances other relevant normative claims of various weights and priorities. Absent competing normative claims, the reasoning process must still take into consideration the potential cost or effort required to bring about the desired state of affairs: A person may value state of affairs *X*, but does he or she value it enough to do everything necessary to bring it about? Even asking such a question presupposes that our time and energy are themselves valuable to us, so there will almost always be competing normative claims at issue when evaluating or justifying prescriptive claims.

So rather than providing immediate *justification* for its parallel prescriptive claim, any given normative claim provides only a *limited partial justification*: A normative claim plays a role in a complex reasoning process wherein it is weighed against the effort required to achieve the valued state of affairs and against any relevant competing normative claims, a process which might or might not lead to a prescriptive claim parallel to the normative claim. In other words, "state of affairs *X* is valuable to *A*" does not necessarily lead to the conclusion "*A* ought to take action to bring about *X*," even if *A* is perfectly capable of the action or actions in question.[7]

However, such complexities are not an obstacle in the analysis at hand, but rather are a strength: Any position that demands too simple and straightforward a bridge between facts and values would poorly reflect the complexity of both ethical theory at an abstract level and practical reasoning (whether ethical or instrumental) at the level of day-to-day decisions. Moreover, this complexity has a direct bearing on one of the key features of ethical theory,

[7] And, naturally, even if the claim "*A* ought to take action to bring about *X*," is true, that will not necessarily lead *A* to take action. It is important in any discussion of prescriptive claims (or any other aspect of ethical theory) not to confuse justification with motivation. It is possible for a prescriptive claim like "*A* ought to take action to bring about *X*" to be true without *A* feeling any motivation or inclination to do so, and vice versa. Even if one insists, as many do, that moral claims have a motivational component – that is, if one asserts that simply believing that a moral claim is true provides the believer with motivation to act as the moral claim demands – it is still the case that motivation is not *the same as* justification. Justification may have considerable bearing on motivation, but it is not identical with it.

which I alluded to but did not attempt to enumerate above: The prescriptive conclusions of any given ethical theory are taken to be in some sense *overriding*. A genuinely moral norm, such as the value placed on human life (by any ethical theory), is generally taken to have greater weight than a non-moral norm in the competition of values leading to a prescriptive conclusion. For example, no matter how highly Bob values financial security, he ought not push his rich elderly uncle off a cliff to collect an inheritance – because the value placed on human life is typically taken to trump any amount of value Bob places on financial security.

Such a coarse example might serve well to illustrate the idea *that* moral values (i.e. moral normative claims) are considered overriding, but that does not necessarily make it obvious *why* moral values override lesser concerns. At first glance, it might seem that even raising the question creates an ugly morass of prescriptive claims about normative claims: "One ought not value financial security more than one values human life" seems easy enough for a start, but what about all the other things one values and disvalues? Why, in particular, should the value of human life trump the value Bob places on financial security? One possible path to an answer lies in the very idea of a *principle*: As I mentioned in the loose sketch of this argument above, an 'ought' statement cannot accurately be called a 'principle' if it is not applicable to everyone. Just as what is of value to an individual provides grounds for reasoning about that individual's life decisions, so what is of value to *every* human would seem to provide at least some grounds for reasoning about *anyone's* decisions. If some state of affairs *X* is in fact of value to every human, then one might have some basis for drawing conclusions about what actions any human should take to best realize state of affairs *X*. In other words, normative claims support prescriptive claims, and if a normative claim can be ascribed to every human being, then the related prescriptive claim would also apply to every human.

However, that reasoning seems only to push the question back a step: Principles are universal prescriptive claims, applying to everyone. But why should a prescriptive claim derived from what is of value to everyone automatically take priority over a prescriptive claim based on what is of value to a given individual? Does the overriding character of ethical principles really just amount to some sort of moral tyranny of the majority? Again, mere agreement does not constitute justification: Determining which values should take priority in any evaluation of competing values should be based on some feature of the valued state of affairs itself, not on the number of people to whom it is valuable. But what feature?

The natural place to begin is to ask why something is valued in the first place. Many of the things we value are only valued for how they contribute to achieving something else. Most of us don't value going to the doctor because it's such a wonderful experience, but we value our health and the occasional visit to the doctor contributes to maintaining and/or restoring our health. This reasoning introduces the traditional distinction between **intrinsic value** (the value/good a state of affairs has in itself) and **extrinsic value** (the value something has only insofar as it advances or contributes to some other valuable state of affairs – good for something else rather than good in itself), which seems to offer at least some of the criteria for weighing competing values discussed above. As a matter of definition, the distinction between the two sorts of value involves a priority claim: Something valued only for its contribution to something else of value is subordinate to the "something else" to which it contributes. Further, it seems plausible that something of intrinsic value should generally be prioritized above that which merely has extrinsic value: At the very least, that which has extrinsic value would seem to carry weight only in proportion to its contribution to a state of affairs of intrinsic value, so we reduce the problem of prioritizing everything valuable to just the problem of prioritizing that which is intrinsically valuable.

The intrinsic/extrinsic distinction seems to offer a basis for both a hierarchy of considerations to be weighed and a basis for judging relative weight in the process of reasoning towards prescriptive conclusions from multiple normative premises. Thus, careful attention to extrinsic and intrinsic value in the process of reasoning towards prescriptive conclusions seems like a plausible way to generate prescriptive claims which have the overriding character generally understood to be a key feature of ethical principles.

There is another common philosophical use of the terms 'extrinsic' and 'intrinsic' which may introduce confusion here. An 'intrinsic property' belongs to an entity in itself, and an 'extrinsic property' belongs to an entity only in relation to the properties of some other entity. For example, the property of being a certain height (roughly 5'9" tall) is intrinsic to me, but the property of being taller than my godson is extrinsic: I have the taller-than property only because my intrinsic property of height stands in a certain relation to his intrinsic property of height. If his current rate of growth is any indication, his intrinsic property of height will soon be such that I will lose that extrinsic taller-than-him property without any change to my own intrinsic property of height.

This dual usage of the terms 'intrinsic' and 'extrinsic' for values and properties can be misleading, and I bring it up precisely to forestall any confusion in this argument: When I say a state of affairs is 'of intrinsic value' or 'intrinsically valuable,' I mean that the state of affairs is valuable in its own right, and not valuable only by virtue of advancing or contributing to some other valuable state of affairs.[8] When I say a state of affairs is 'of extrinsic value' or 'extrinsically

[8] Note that the inclusion of "only" permits a state of affairs to be both intrinsically valuable and extrinsically valuable. That is, a given state of affairs might be valuable in its own right and also valuable

valuable,' I mean that the state of affairs is valuable by virtue of advancing or contributing to some other valuable state of affairs. This does not in any way imply that being valuable is an intrinsic *property* of any state of affairs, for value is always a relational concept: Any state of affairs can be properly said to have value if and only if there is some entity for which it has value. Value always implies a valuer, and so in that sense value is always extrinsic/relational, based on a connection between a valuing entity and a valued state of affairs.[9]

Note also that I have been careful to refer to the value (or disvalue) of states of affairs and never the value of an object. To say that an object (or class of objects) *O* is valuable is a colloquial usage which acts as a short-hand for "state of affairs *X* is valuable to *A*" in those cases where *X* is something like the following: "*A* possesses *O*," or "*A* has access to *O*." An object does not itself have value, but rather what has value is some state of affairs where some valuer has possession of or access to that object. Objects have (intrinsic or extrinsic) properties, but not (intrinsic or extrinsic) value. States of affairs, in contrast, can be valuable or disvaluable (or value-neutral, of course).[10]

for advancing or contributing to another valuable state of affairs. A state of affairs is *purely extrinsically valuable* only when its sole and entire value depends on its advancement of or contribution to some other valued state of affairs. Similarly, a state of affairs is *purely intrinsically valuable* when it cannot be said to contribute to or advance any other valuable state of affairs, and thus is only valuable in its own right.

[9] Some philosophers do seem to use "value" differently, such that a state of affairs could have value in itself without regard to or relation with any valuer – which might be termed "inherent value," as distinct from "intrinsic value." While I find such a conception of value incoherent on the face of it, it would be a distraction from the matter at hand to categorize and argue against such conceptions: Most conceptions of value discussed in the history of ethical theory are consistent with the account I give here, and in any case the reader may take it as stipulated that when I use the word "value" and its cognates, I refer to a relational concept as I've specified here.

[10] Whether or how states of affairs have properties, or whether "value" is properly conceived of as a "property" at all under appropriately careful definitions, is not relevant to my argument in any way. All that matters here is (1) being of value is not a property *of objects*, and that if it were a property of states of affairs it would be extrinsic *as a property*, requiring a certain relationship between a valuer and said state of affairs, and (2) the usage of 'extrinsic' and 'intrinsic' with respect to value is simply not the same as (and must not be confused with) the usage with respect to properties.

However, it is not only terminology that causes confusion when it comes to intrinsic and extrinsic value. It is quite possible, and indeed common, for any given human to be confused about what is extrinsically valuable and what is intrinsically valuable. For example, some people place so much importance on acquiring wealth that they quite forget that money is only valuable for what it can buy, and there is much of value that money cannot buy: Perhaps most famously, money can't buy you love.

While citing such pop music sentiments may seem flip, it is difficult to argue against the observation that many people have undermined relationships with spouses, children and other loved ones by devoting a disproportionate amount of their time and energy to the acquisition of wealth, much to the detriment of their own well-being and that of their loved ones. And however 'well-being' is defined, it would seem to be something of intrinsic value: It is extremely implausible that anyone values their own well-being solely for the sake of something else.[11] In contrast, the extrinsic value of money would seem at least partially based on how it contributes to one's well-being, since money is by definition something which has value primarily or solely because it is exchangeable for other things, some of which can be basic contributions to well-being such as food, shelter, and so on (and potentially more elaborate contributions to well-being, such as funding an education). To prioritize something of extrinsic value above the intrinsically valuable states of affairs to which it contributes is clearly a mistake.

Even such a cursory reflection on everyday human experience with regard to value thus reveals that value is something about which we can be mistaken. We can value incorrectly, such that we prioritize the less valuable over the more valuable, and potentially even place value on

[11] Although one might prioritize someone else's well-being (perhaps that of one's child) *above* one's own, that is not quite the same as treating one's own well-being as purely extrinsic (i.e. valuable only insofar as it contributes to the well-being of one's child).

(or mistake as valuable) that which actually has negative value (disvalue). In other words, we can think that some thing is good without it actually being good. That this possibility is immediately understandable and widely recognized indicates that, at least on the ordinary understanding of the term, 'value' is sometimes taken to be non-arbitrary: We can and do acknowledge a difference between that which is truly valuable to us and that which we merely *think* is valuable, even if many of us sometimes fail to remember and draw this distinction. In other words, this distinction recognizes the verb 'value' and the noun 'value' as being potentially disjoint: People can and do value states of affairs that do not actually have value for them, and possibly even those which have negative value (or disvalue). This is why I have been careful throughout this argument to distinguish use of the noun and adjective forms "of value to *A*" and "valuable to *A*" from the transitive verb "*A* values..." The former can at least potentially be interpreted in such a way that the matter of value is separate from *A*'s personal opinions or desires, but the latter is much more naturally interpreted as a statement primarily about some cognitive state of *A*'s.

This insight about mistakes in valuing draws a distinction between what has merely **subjective value** and that which has **objective value**. While the observation may seem somewhat pedestrian, in the context of ethical theory it constitutes a substantive and controversial position. In fact, I am taking a position on one of the metaethical disputes I mentioned in my introduction (see p.8). The notion that value is something about which we can be mistaken implies some version of 'moral realism' – depending on how one defines that notoriously slippery phrase. However, the various definitions of and arguments for and against 'moral realism' and 'moral anti-realism' in the history of metaethics would be a distraction from the argument at hand. Instead, I will simply clarify the definition for 'objective value' that I am using here: Some moral value claims – not necessarily all value claims, nor even all moral value claims – are false, and others are true. This still allows for the possibility that other moral value claims – perhaps even all moral value claims but one foundational claim – are neither true nor false, and perhaps are not analyzable in terms of truth or falsehood at all. Furthermore, I am not stipulating or specifying what in the world determines the truth or falsity of any particular moral value claim, nor suggesting any particular method for determining the truth or falsity of any given moral value claim. Such details will depend on a particular account of what is of objective value and why, so cannot be addressed in the abstract.

Note, however, that I have not actually argued that moral realism (by whatever definition) is definitively true: I have only claimed that it is a widely held view that we can be mistaken about matters of value, and grounded that view with some plausible observations from broadly shared human experience. This is not a particularly strong argument – nor is it intended to be. Without some much richer and more specific account of what is of objective value and why, no stronger argument can yet be made. For now, the plausible observations on which the distinction between subjective and objective value has been made will serve as a kind of placeholder, to be filled in by the detailed arguments for any given ethical theory.

But even in the absence of a more substantive argument for moral realism, the distinction between objective and subjective value plays an important role in the search for a solution to the fact-value problem. Hume's problem is a matter of justification, and justification is generally taken to require more than providing personal opinion or individual perception or something else subjective. In the context of justification, the acknowledgement that it is possible to be mistaken about some category of claims comes with a presumption that someone making a claim of that sort will provide reasons to believe the claim is true. Thus, the possibility of justifying moral claims implies the existence of moral truths: In the absence of any underlying truth for its foundational value claims, there would be no reason to view the prescriptive claims of any ethical theory as justified; 'ought' claims would not be particularly compelling, or even interesting, let alone obligatory.

Justification thus demands a strong distinction between appearance and reality, between matters of opinion and matters of fact. The subjective/objective value distinction implies that value need not merely be a matter of preference or opinion, even in the absence of any proposed way to determine what is objectively valuable rather than merely subjectively valued. A given value claim need not refer solely to some psychological state wherein a person believes or perceives that a given state of affairs is valuable, or feels some sort of desire or motivation to bring about that state of affairs: Instead, a value claim might refer to a relation that obtains in the world wherein a given state of affairs is of value (or disvalue) to that person without regard to any psychological state or inclination that person may have.

In other words, whether a given person desires some state of affairs or *perceives* it as valuable does not determine whether or not it *is* valuable. This implies a particularly strong sense of 'objective' at work in this conception of 'objective value,' more than mere consensus among subjective impressions or beliefs. A state of affairs could be of objective value to every human who ever was or will be without depending in any way on the intentions, motivations, desires, knowledge, or even awareness of any human. For example, one could intelligibly claim that gaseous oxygen was of value to humans (and all aerobic organisms, for that matter) long before any human knew what an element or molecule was, or that O_2 is a major component of Earth's atmosphere.[12]

[12] More precisely, a state of affairs where the partial pressure of gaseous O_2 falls within a given range is of objective value to humans. Different aerobic organisms can tolerate different partial pressures, and some can extract O_2 dissolved in water in varying ranges of concentrations. More importantly, some subpopulations of humans whose ancestors have lived at high altitudes for many generations can withstand lower partial pressures of oxygen without serious detriment, and might even suffer some ill

Taken together, the considerations and distinctions developed in this section say about as much about what sort of foundational value claim might successfully bridge the fact-value gap as I think can be said without actually proposing a specific candidate claim. The identification of some state of affairs which has intrinsic, objective value for each and every human would seem to serve as a basic normative premise from which prescriptive conclusions could be justifiably derived, potentially culminating in the universal prescriptive claim(s) of a complete ethical theory. Identification of the normative claim as being true for everyone creates the possibility of arguing towards not just prescriptive claims, but **universal** prescriptive claims which are true for everyone. Identification of the normative claim as **intrinsic** makes it possible to determine the weight of various potential normative claims in order to arrive at those universal prescriptive claim conclusions, as well as reflecting the common position within the tradition of ethics that ethical precepts override prescriptive claims derived from lesser concerns, e.g. other normative claims which are either extrinsic or subjective, possibly both. Finally, identification of the value claim in question as **objective**, i.e. factual, would answer the primary Humean concern that the relation between fact and value "shou'd be observ'd and explain'd." What remains is that "a reason should be given," or that **justification** be offered for some particular foundational value claim or claims.

It is important to be clear how much is at stake for ethical theory in this proposed solution to the fact-value problem. First, arguments must be made to establish that some state or states of affairs are as a matter of fact (i.e. objectively) of intrinsic value to each and every human. Establishing such facts about what is intrinsically valuable to humans forms the bridge across

effects at sea level. I raise these complicating details here simply to emphasize the fundamentally pragmatic foundations of this conception of value relations: Statements like "oxygen is of value" must always be taken broadly to indicate a state of affairs where oxygen is present in whatever concentration and form makes it available for metabolism by a given aerobic organism is of objective value to that organism (and relevantly similar organisms).

Hume's fact-value gap, insofar as such an argument constitutes justification for at least one value claim on non-value grounds. Without at least one such value premise, no value conclusions can be supported – including the universal prescriptive claims of any ethical theory, such as Mill's version of the principle of utility or Kant's categorical imperative. Thus, although many further arguments might be necessary to connect any given fact-value bridging normative claim(s) to the universal prescriptive claim(s) of an ethical theory, such an initial value claim premise is a logically necessary step and can rightly be described as the **foundation** for any subsequent ethical theory: The argument for any given ethical theory's ultimate prescriptive principle or principles can only be as well-justified as the argument which justifies some normative claim(s) on non-value grounds, i.e. the argument which establishes some fact-value bridging claim.

In the remaining sections of this chapter, I will argue that three approaches to ethical theory traditionally identified as central – utilitarianism, deontology and virtue theory – are all structured around such foundations: For example, J. S. Mill grounds utilitarianism in claims about human psychology, arguing that happiness, defined as pleasure and the avoidance of pain, is the only thing valued for its own sake by every human. Kant grounds his deontological ethics in claims about *a priori* necessity, arguing that every rational being as a matter of logical necessity values the rational will that gives the moral law unto itself. Aristotle grounds his virtue ethics in claims about natural function, arguing that a flourishing human life governed by reason is the highest good towards which all our actions ought to aim.

In each of these approaches to ethical theory, arguments are advanced to justify the claim that some state of affairs *X* (happiness, exercise of the rational will, human flourishing) is as a matter of fact intrinsically valuable to every human. These arguments play the exact role proposed above as a plausible solution for Hume's fact-value problem. Without these fact-value-bridging claims about human nature to serve as value premises justified on non-value grounds, no further value conclusions could be drawn – and therefore no ethical theory could be advanced.

Whether examining the foundations of traditional ethical theories or advancing my own foundation (as I will in Chapters 3 and 4), framing this project in terms of solving Hume's problem has set up a conditional argument: If it is possible for a given prescriptive ethical theory to be justified, then that theory must at the bare minimum provide a solution to Hume's problem by identifying and justifying a fact-value bridging claim that meets the criteria outline above. (I say "at a bare minimum" because further arguments may be – and usually are – required to connect the foundational fact-value bridging claim to a complete prescriptive ethical theory.) I call this a *conditional* rather than *transcendental* argument because a the latter requires some basis, something that serves as a given. In a transcendental argument, some phenomenon or array of phenomena (having certain experiences, for example) serves as the accepted premise upon which a conclusion can be based: If there are some conditions without which the phenomenon would not be possible, and the phenomenon is not only possible but actual, then those conditions for the possibility of the phenomenon must obtain. While I have outlined the conditions for the possibility of justifying an ethical theory, no extant ethical theory is an already-accepted premise which can serve as the foundation for a transcendental argument establishing that those conditions for justifying an ethical theory must obtain.

To be clear, I am not taking it as a given that justification is possible for *any* ethical theory. In this section, I have in effect presumed that what I've labeled Hume's problem can be solved, and outline how it can be done *given that presumption*. In effect, the criteria I developed for a fact-value bridging claim – objectivity, universality, etc. – all depend on the presumption that normative value claims can be rigorously justified: However, no argument I have made so far has justified that presumption, or even pretended to do so.

What, then, will I have accomplished even if I do succeed in developing a fact-value bridging claim about human nature justified on evolutionary grounds? The answer hinges on one's opinion regarding the antecedent of my conditional argument: If one believes it is possible to justify a prescriptive ethical theory at all, then it is a very significant accomplishment to generate a fact-value bridging claim about human nature based on evolutionary biology which can serve as the foundation for a prescriptive ethical theory. It is especially significant in light of fact that many philosophers claim that a prescriptive evolutionary ethical theory is impossible. Moreover, the concerns that motivate that claim of impossibility are, for most philosophers, exactly the same concerns that Hume raises – that is, the problem of drawing value conclusions from factual premises, broadly construed. Framing my argument around solving Hume's problem directly addresses those concerns.

If one rejects the very possibility of providing genuine justification for any prescriptive ethical theory, it might seem as if my overall argument in this dissertation has nothing to offer. I do not think this is quite so obviously the case, however. At the risk of appearing to affirm the consequent, successfully advancing and defending a foundation for a prescriptive evolutionary ethical theory appears to offer some hope for the possibility of justifying prescriptive ethics. While I have already emphasized that it is not a given that justification of any ethical theory is possible, it is also not a given that justification of any ethical theory is *impossible*. If the question is open – and the tangled, unsettled nature of metaethical debate I alluded to in my introduction (see p.8) is a strong indicator that the question is very open indeed – then a great deal hinges on why a given philosopher rejects the possibility of justification. One reason to be skeptical that justification is possible is simply that one is unconvinced by any of the justifications that have ever been offered: Even if the perceived failure of prior attempts at justification is not a direct premise for a given metaethical position, it often seems to be the motivation for delving into many kinds metaethical questions in the first place. If I succeed in offering a more convincing justification for the foundation of ethical theory than has been offered by the traditional theories I am about to examine, I will have undermined one of the main reasons for widespread skepticism about the possibility of justifying prescriptive moral claims.

Section 3: Utilitarianism via John Stuart Mill, with refinements of solution

Perhaps the clearest example of an ethical theory built upon a foundational fact-value bridging claim as described above is utilitarianism. In Chapter IV of *Utilitarianism*, titled "Of What Sort of Proof the Principle of Utility Is Susceptible," John Stuart Mill begins thus:

> It has already been remarked, that questions of ultimate ends do not admit of proof, in the ordinary acceptation of the term. To be incapable of proof by reasoning is common to all first principles; to the first premises of our knowledge, as well as to those of our conduct. But the former, being matters of fact, may be the subject of a direct appeal to the faculties which judge of fact – namely, our senses, and our internal consciousness. Can an appeal be made to the same faculties on questions of practical ends? Or by what other faculty is cognisance taken of them?
>
> Questions about ends are, in other words, questions what things are desirable. The utilitarian doctrine is, that happiness is desirable, and the only thing desirable, as an end; all other things being only desirable as means to that end. What ought to be required of this doctrine – what conditions is it requisite that the doctrine should fulfil – to make good its claim to be believed?
>
> The only proof capable of being given that an object is visible, is that people actually see it. The only proof that a sound is audible, is that people hear it: and so of the other sources of our experience. In like manner, I apprehend, the sole evidence it is possible to produce that anything is desirable, is that people do actually desire it. If the end which the utilitarian doctrine proposes to itself were not, in theory and in practice, acknowledged to be an end, nothing could ever convince any person that it was so. No reason can be given why the general happiness is desirable, except that each person, so far as he believes it to be attainable, desires his own happiness. This, however, being a fact, we have not

> only all the proof which the case admits of, but all which it is possible to require, that happiness is a good: that each person's happiness is a good to that person, and the general happiness, therefore, a good to the aggregate of all persons... [Mill, p.269]

Before addressing Mill's argument, I want to focus specifically on his use of a fact-value bridging foundational claim as I have characterized it in the previous section. Happiness, which Mill has minimally defined in prior chapters as pleasure and the avoidance of pain [Mill, p.239], is of value to each and every person according to the above. Further, Mill spends the remainder of Chapter IV arguing that happiness is not only valued as an end in itself as a matter of psychological fact, he also argues that it is the only thing valued solely as an end in itself. Thus, happiness qualifies as a state of affairs that is of intrinsic, objective, universal value. The value of happiness serves as the normative foundation for reasoning towards the universal prescriptive claim conclusions of utilitarianism as an ethical theory, and Mill explicitly acknowledges that it plays this foundational role both in how he advances his argument and by calling it a "first principle."

One might object to the description of happiness as *objectively* valuable in this context. After all, what people find to be desirable is not only generally understood to be subjective, but is perhaps the paradigm realm of subjectivity. The subjectivity of desire is so established that it's enshrined in the Roman aphorism that there is no disputing matters of taste, *de gustibus non est disputandum*. Would it even make sense to say that a person could be mistaken about what he or she takes pleasure in?

However, what seems most relevant here is that the value of happiness itself is taken by Mill to be a matter of plain fact: If some people believe that they do not value happiness as an end in itself, or even claim that they value something else as an end in itself without also valuing it as a means to achieving happiness, Mill would contend that such people are simply mistaken.

Great variety in what activities individuals find pleasurable (and painful) is acknowledged by Mill; he not only acknowledges it, he even cautions that it is important for the general good that society attempt to educate people so that they take pleasure in the right activities, such as advancing the general good. But no amount of variation in what people find pleasurable changes the fact that all people value pleasure (and the absence of pain) as an end in itself.

Such a response, however, does not quite answer the objection. Even if it is true that all people value happiness as an end in itself, surely what each values is *his or her own* happiness as an end in itself, which seems awfully subjective on the face of it. In that light, Mill's argument as quoted above seems highly problematic. How can Mill move from the evidence that people do as a matter of fact desire happiness to the conclusion that happiness is desirable in and of itself – and therefore *ought* to be desired – without committing a rather blatant is-ought fallacy? And how does the fact that each person desires *their own* happiness in any way support the conclusion that the *general* happiness is desirable? And if the only proof of the desirability of something is that it is desired, by whom is this general happiness desired? The evidence Mill puts forward is that each person desires his or her own happiness, not that each person desires the general happiness – whatever that is! To resolve these problems, it is necessary to understand what kind of evidence Mill sees our desiring to be, and of what precisely it serves as evidence.

Consider Mill's comparison of desirability to visibility and audibility in the paragraphs quoted above: Is seeing something really *proof* of its visibility? An individual person's claim to have seen or heard some particular object or event surely doesn't constitute absolute proof that the object or event is visible or audible, because people are subject to errors, hallucinations, dreams, confusion and so on. Nevertheless, simply by virtue of what the words 'visible' and 'audible' mean, the only sort of evidence we ever do have or could have for the visibility or audibility of any given object or event is people seeing or hearing it. Remember, Mill's contention is not that perceiving is proof of visibility or audibility, but that it is "[t]he only proof capable of being given" – that is, the only sort of evidence we can ever have. Similarly, Mill's analogy suggests that a given person desiring some particular thing is not necessarily proof that this particular thing is desirable, but it remains true in a general sense that the only kind of evidence we ever do have or could have of the desirability of something is people desiring it.

Geoffrey Sayre-McCord addresses many of these same concerns about Mill's argument in an essay titled "Mill's 'Proof' of the Principle of Utility: A More than Half-Hearted Defense," and he is particularly clear in explaining this distinction:

> ... In the case of our senses, the evidence we have for our judgments concerning sensible qualities traces back to what is sensed, to the content of our sense-experience. Likewise, Mill is suggesting, in the case of value, the evidence we have for our judgments concerning value traces back to what is desired, to the content of our desires. Ultimately, the grounds we have for holding the principles we do must, he thinks, be traced back to our experience, to our senses and desires. Yet the evidence we have is not that we are sensing or desiring something but what it is that is sensed or desired.
>
> When we are having sensations of red, when what we are looking at appears red to us, we have evidence (albeit overrideable and defeasible evidence) that the thing is red. Moreover, if things never looked red to us, we could never get evidence that things were red, and would indeed never have developed the concept of redness. Similarly, when we are desiring things, when what we are considering appears good to us, we have evidence (albeit overrideable and defeasible evidence) that the thing is good. Moreover, if we never desired things, we could never get evidence that things were good, and would indeed never have developed the concept of value.
>
> ... "Desiring a thing" and "thinking of it as desirable (unless for the sake of its consequences)" are treated by Mill as one and the same, just as seeing a thing as red and thinking of it as red are one and the same. Accordingly, a person who desires **x** is a person who *ipso facto* sees **x** as desirable. Desiring something, for Mill, is a matter of seeing it under the guise of the good. This means that it is important, in the context of Mill's argument, that one not think of desires as mere preferences or as just any sort of motive. They constitute, according to Mill, a distinctive subclass of our motivational states, and are distinguished (at least in part) by their evaluative content. Thus, Mill is neither assuming nor arguing that something is good because we desire it; rather, he is depending on our desiring it as establishing that we see it as good. [Sayre-McCord, p.339-340]

Perhaps the claim that desirability is a property which can only be inferred from our desires does not completely dispel the is-ought fallacy concern, but the plausibility of the position does seem to shift the burden of proof to a critic. That is, someone who is not satisfied that desirability can only be inferred from desires would seem to be obligated to suggest what other evidence we could or should have before we can draw conclusions about whether something is desirable.[13] Alternately, a critic could argue that desirability is not really a property of anything at all, regardless of the evidence of our desires.[14] Given the universality of the desire for happiness, it would seem that the critic would also need to provide either a counterargument to Mill's claim that happiness is universally desired, or some explanation why everyone is mistaken in desiring happiness as an end in itself.

Consider also that, despite the fact that the only evidence for visibility or audibility we have is based on our experiences, the visibility of an object or audibility of an event is nevertheless an objective property of the object or event. Imagine that I claim that some particular thing is visible in front of me right now. Imagine that you are standing right next to me and fail to see whatever it is I'm claiming is visible, and tell me so. If I answered, "That doesn't matter. It's visible because *I* see it, whether you see it or not," you could justly claim that I have quite failed to understand what the word 'visible' means. A thing that is visible-for-me and invisible-for-you is not by any reasonable definition 'visible' as such, without qualification. While determining whether an object has the property of visibility does depend on experience, it is not determined by the occurrence or nature of any particular person's observation; visibility is

[13] Kant, for example, has a rather different picture of both what is good and how we can have knowledge of the good, which will be discussed below.

[14] For example, J.L. Mackie argued that all moral value claims – such as the claim that happiness is desirable/good – are mistaken because values do not exist in the world and thus cannot be referred to. This position, which Mackie dubbed "error theory," expresses but one possibility within the metaethical debates I alluded to in my introduction, and counts as one species of moral anti-realism. [Mackie, 1977]

not a subjective property obtaining only in some relation between an object and a given person. Rather, visibility is a property of the object, and only if an object can be seen by anyone and everyone who happens to come along can we sensibly say that seeing it constitutes evidence for the property of visibility. Mill's analogy suggests that the same is true of desirability. But what, exactly, does this imply when applied to desirability?

If desirability is truly like visibility or audibility, it must be an objective property – happiness is desirable without qualification, not simply desirable-for-me from the perspective of any particular person. Even though determining whether something has the property of desirability does depend on experience, it is not determined by the occurrence or nature of any particular person's desiring, so it is not a subjective property obtaining only in some relation between happiness and a given person: To say that one and the same thing – e.g. happiness – is desirable-for-me but undesirable-for-you would indicate a failure to understand what the word 'desirable' means.

But isn't that the very question that is at issue? Whether or not happiness is an objective property of this sort as opposed to a subjective one – each person desires their own happiness, after all – is the question at hand: If the analogy simply assumes that desirability must be an objective property like visibility and audibility, it does little to establish the point.

However, I think the analogy still advances the argument if one considers the role of evidence in the analogy carefully. The failure to understand 'visibility' in the hypothetical example offered above depends primarily on a failure to understand how (and to what extent) my perception constitutes evidence for visibility: In the example, my error was to take my seeing an object as evidence of its visibility *without regard to whether others would or could also see it were they in my place*. If I understood the concept of visibility properly, I would know that an object having the property of visibility means that anyone can see it if he or she is in a position to do so, and I would not say something so absurd as "It's visible because *I* see it, whether you see it or not."

This does not entirely deny the existence of more subjective, relational sorts of properties. With respect to properties other than visibility, it's easy to imagine taking our perceptions to be evidence for something much more subjective. That is, it might be quite reasonable for me to conclude that an item of clothing has that property of being "too brightly colored" or "the wrong shade for me" without regard to whether anyone else would see the same thing in my place. But for most properties we attribute to objects – whether specific properties like having a particular shape and color or general properties like visibility – in thinking that our perceptions provide evidence that the object has the property, we also think that anyone else in our position (with adequate sensory capacity, appropriate lighting, line of sight, etc.) would see more or less the same thing.

Similarly, perceiving a state of affairs to be desirable carries with the same sort of understanding that anyone in a relevantly similar position would perceive its desirability. There could also be idiosyncratic desires peculiar to an individual, some things I might desire specifically for myself without thinking anyone else would necessarily find them desirable. But for the most part when we perceive some state of affairs as desirable, we not only think it is good to achieve that state of affairs for ourselves, we think that anyone else who achieves that state of affairs will have gotten something good. Again, Sayre-McCord offers a particularly cogent summary of this idea:

> ... [I]magine a rich, luscious, moist, chocolate cake (which, I will be assuming, is a lot like happiness, at least in its being desirable)... [W]hatever it is about the cake that I am desiring, and so desiring for myself, is something the cake may still have if someone else were to get it. In seeing the cake as valuable

> (and so worth getting for myself), I therefore seem to be committed to thinking that if you should get it rather than me, you have gotten something good. Of course, I am not committed to desiring that you get it, nor to thinking that your getting it is a good to me. However, the grounds I have for thinking it would be a good to me, were I to get it, appear to commit me to thinking that you would be getting something good were you to get it.
>
> It is important here to keep in mind the contrast between my desiring the pleasure I might get from the cake and my desiring the cake. If it is the pleasure I desire (for myself), then my commitment *vis-à-vis* you would be to seeing your getting a similar pleasure as your getting something good. If the cake does not bring you that pleasure, then in getting a piece of the cake you would not be getting what I take to be good about it, and the difference could consistently be seen by me as making all the difference. What I cannot consistently do is see as valuable some feature of the cake, value getting the cake for myself on those grounds, and then deny that when you get it (with that feature), you get something good.
>
> Analogously, if each of us is, in desiring happiness, desiring not merely *our own* happiness, but desiring happiness (for some nonproprietary feature of it) for ourselves, we cannot consistently then deny that when someone else gets happiness (with that feature), they get something good. Of course, again, we are not committed to desiring that someone else get it, nor to thinking that their getting it is a good to us. Nevertheless, the grounds we each have for thinking it would be a good to us, were we to get it, appear to commit us each to thinking that in getting it, someone else would be getting something good. [Sayre-McCord, p.347-348]

So it is not quite correct to say that people only value their own happiness. Rather, people value happiness – that is, they see happiness as objectively desirable – *and* they want it for themselves. The motivation may be – and generally is – quite proprietary in that one wants it for oneself, but to desire something is to perceive it as desirable and therefore want it, not simply to perceive it as desirable-for-me (but not for anyone else).

This does not require any strange view that people must take pleasure in the same sorts of things. Rather, it simply requires admitting that when someone else achieves something pleasurable – with full acknowledgment that other people can and often do take pleasure in very different sorts of experiences and activities than oneself – they are getting something good. In fact, we could not make sense of the common phenomenon of taking pleasure in other people's pleasure – even something so universal as parents' delight in seeing their children happy – without understanding the value of pleasure in this non-proprietary way. Why would one care about a friend's happiness if one perceived happiness only as good-for-me-to-get rather than simply good, and thus good for one's friend to get as well?

Realizing that value/desirability of happiness has this non-proprietary[15] character is the key to unlocking the other mysteries of Mill's argument as well. So how does the fact that each person desires *his or her own* happiness in any way support the conclusion that the *general* happiness is desirable? People may be most strongly motivated to pursue their own happiness, but what desires give evidence for is the value of happiness itself, no matter whose. Given the objective, non-proprietary desirability of happiness – that is, given that happiness is simply good, rather than good-for-me – more happiness is better without regard to whose happiness it is. Thus it becomes possible to say that the general happiness is desirable, insofar as the phrase "general happiness" means no more than the aggregation or summation of individuals' happiness.

And if the only proof of the desirability of something is that it is desired, by whom is this general happiness desired? Mill does not assume that we are necessarily motivated by the general happiness: In fact, he specifically argues that we can and should *teach* people to take pleasure in promoting the general happiness in order to enhance their motivation to do good. [Mill, p.260-268] Rather, our desires only give evidence for the objective, non-proprietary desirability of happiness. The aggregate happiness is not something that is desired for its own sake, but is a logical consequence of the non-proprietary desirability of everyone's individual happiness. We do not necessarily value the general happiness as individuals, but the non-proprietary value of

[15] Although Sayre-McCord chooses not to hyphenate "nonproprietary," I have chosen to use "non-proprietary" instead to avoid terminological confusion: The non-hyphenated word already has an established definition – an adjective for that which is not protected by trademark, patent or copyright.

happiness which we infer from our individual desires is the basis (the only possible and available basis, Mill argues) for concluding that the general happiness is also valuable. The value of happiness for every individual may be a matter of human psychological fact, but the principle of utility itself is an achievement of reason, not natural inclination.

This matter of non-proprietary value was raised while resolving the surface appearance of certain logical problems in Mill's particular argument, but it seems to be very close to the heart of what a successful fact-value bridging normative claim must be. If so, why didn't non-proprietary value come up in the previous section as one of the criteria a satisfactory fact-value bridging claim must meet? As I will discuss below, non-proprietary value is a necessary consequence of any value being both objective and universal; but when I originally discussed the importance of objective value, I had already eliminated non-universal value claims from consideration, so the matter simply didn't come up.

As a useful example, let us consider self-interest: Self-interest seems to have a certain ineluctably proprietary nature, but that doesn't mean it is subjective in the sense that I have defined the term. It seems quite reasonable to say that some state of affairs might as a matter of fact be in a given person's best interest regardless of whether the person is even aware of its value, or even if the person actively denies its value – that is, self-interest might plausibly qualify as objectively valuable. But even if objectively valuable in this sense, a state of affairs which is valuable as a matter of pure self-interest is still proprietary: That is, a state of affairs that benefits my self-interest is valuable-to-me, not simply valuable from the perspective of anyone and then desired by me as such. So it is possible for some state of affairs to be of objective, proprietary value.[16]

[16] One problem to be aware of and cautious about here is different 'subjective/objective' distinctions which are easy to confuse or conflate. We use 'objective' to distinguish matters of fact from matters of

Yet, self-interest is generally taken to be universal in some sense. Every person values his or her own self-interest, right? Not so fast. If what is objectively valuable to a given person is valuable to that person in a strictly proprietary fashion, then it cannot also be a universally valuable state of affairs: Rather, some state of affairs *X1* (the realization of some state of affairs for some person designated 1) is objectively valuable to *P1* (some person designated 1), and an otherwise identical state of affairs *X2* is objectively valuable to *P2*, but there is no state of affairs *X* which is valuable to both *P1* and *P2*, let alone to them and everyone else in the world.

That's awfully abstract, so to clarify let's return to the Sayre-McCord's chocolate cake analogy, ignoring for the moment that chocolate cake is not really objectively valuable (certainly not to a diabetic, for example). Suppose there's not enough chocolate cake to go around: Suddenly, tensions arise between non-proprietary and proprietary value. In perceiving a feature of the cake to be desirable/good, I am still committed to realizing that when you get cake (with that feature), you are getting something good. But I also perceive getting a piece of cake *for myself* as desirable, and when there isn't enough cake to go around, your getting cake becomes undesirable to me because it prevents my getting cake, and vice versa. State of affairs *X* (eating yummy chocolate cake) may be universally valuable, but state of affairs *X1* (*P1* eating cake) is valuable only to *P1* and would have negative value for every cake-lover other than *P1* if cake were scarce. State of affairs *X1* is valued in a proprietary fashion (it's good for *P1* to eat chocolate cake), but *X1* is not valuable to just anyone in the way state of affairs *X* is valuable (it's good to eat chocolate cake), so such a proprietary value cannot also be universal.

mere 'subjective' opinion. But we also sometimes use the word 'subjective' to distinguish claims whose truth conditions vary depending on who utters them from claims which are true regardless of who utters them: When I write or say "I prefer sorbet to ice cream," it is a true claim about the world, whereas it would be false if my roommate said the same words, and so it is 'subjective' in this sense. In contrast, a sentence like "George prefers sorbet to ice cream" remains true regardless of who utters it, and is thus 'objective' in this sense. I am not using objective/subjective terminology in this truth-conditions sense anywhere in this argument.

Now let's set aside cake and return to self-interest. There might be some abstract, content-free sense in which we can say that self-interest is of universal value; we might think that it is of objective value for a person to realize a state of affairs in his or her self-interest regardless of who the person is. But such an interpretation no longer truly addresses *self*-interest, instead placing objective value on interest satisfaction without regard to whose interest is satisfied – and therefore is non-proprietary. Alternatively, we can simply ignore the contradictions which arise between my self-interest and your self-interest and simply accept that some state of affairs (like *X1* above) is valuable to me and disvaluable to you. But if we allow such direct contradictions, we are necessarily denying that the value of such a state of affairs is a matter of objective fact rather than subjective, perspective-limited opinion.

Thus have all the alternatives been considered: If the value of self-interest (or anything else) is objective and proprietary, it cannot be universal. If the value is universal and objective, it cannot be proprietary. If it is universal and proprietary, it cannot be objective. Therefore, any state of affairs which has both objective and universal value must be valuable in a non-proprietary sense as well.

Since it took such careful analysis to show how non-proprietary value is necessitated by the combination of objective and universal value, and since the matter of proprietary vs. non-proprietary value will be very important in subsequent chapters, I will hereafter treat 'non-proprietary' as a separate criterion for the sort of foundational value claim capable of bridging the fact-value gap. In other words, although non-proprietary value may be logically implied by any conjunction of objective and universal value, it appears to be very important for the construction of ethical theory and therefore should be explicitly addressed rather than merely implied.

Having established the importance of non-proprietary value, I will return to the main line of my argument. My contention is that utilitarianism relies on a fact-value bridging normative claim of the sort I characterized in the previous section, and further that this claim plays a central – indeed, foundational – role in the arguments that support the principle of utility. If Mill's argument remains unconvincing even after the explanation and defense presented here, that in no way detracts from the larger argument that such a fact-value bridging normative claim lies at the heart of Mill's ethical theory. If a reader is not convinced by Mill's argument that happiness (defined as pleasure and the absence of pain) is the one and only thing of intrinsic, objective, non-proprietary value to all persons, or is not convinced by the arguments that lead from that foundational claim to the principle of utility (which I have not even discussed in any great detail), then so be it. Frankly, I'm not at all convinced myself.

Note, however, that I have in no way claimed that any philosophers who have advanced any ethical theories have presented airtight justifications for any of their various fact-value bridging foundational normative claims. I have not even claimed that they offer somewhat convincing justifications.[17] I have only contended that establishing the existence of such a fact-value bridging foundational normative claim is *a* solution to Hume's problem. (Perhaps it is the only possible solution, but that's a more difficult claim that I've chosen not to argue here.) Further, I contend that careful examination of the ethical theories advanced by Mill, Kant and Aristotle – as representatives of utilitarianism, deontology, and virtue ethics respectively – will reveal that they do all rely on a fact-value bridging normative claim as I have characterized it. Mill does so very clearly, so I will proceed to Kant.

[17] Of course, if any of these ethical theories had such convincingly justified foundations – either in my own opinion or the collective opinion of the philosophical community – then this paper's attempt to look for firmer foundations in an evolutionary understanding of human nature would be completely unmotivated.

Section 4: Deontology via Immanuel Kant

In a way, Kant's first major work on ethical theory, *Groundwork for the Metaphysics of Morals*, begins in the same place I began this chapter. Kant titles his first chapter, "Passage from the Common Rational Knowledge of Morality to the Philosophical" [Kant, p.195 (4:393)] for much the same reason I started looking for a solution to Hume's fact-value problem by asking what a solution might look like (see p.16): Without a destination, it's difficult to plan a route. That destination is set by the very first sentence of *Groundwork*, Chapter One, which reads: "It is impossible to imagine anything at all in the world, or even beyond it, that can be called good without qualification – except a *good will*."[Kant, p.195 (4:393)]

After a few paragraphs of very broad-brush arguments in support of this strong claim, Kant clarifies the claim and strengthens it still further:

> A good will is not good because of its effects or accomplishments, and not because of its adequacy to achieve any proposed end: it is good only by virtue of its willing – that is, it is good in itself. Considered in itself it is to be treasured as incomparably higher than anything it could ever bring about merely in order to satisfy some inclination or, if you like, the sum total of all inclinations. Even if it were to happen that, because of some particularly unfortunate fate or the miserly bequest of a step-motherly nature, this will were completely powerless to carry out its aims; if even with the utmost effort it still accomplished nothing, so that only good will itself remained (not, of course, as a mere wish, but as the summoning of every means in our power), even then it would still, like a jewel, glisten in its own right, as something that has its full worth in itself. Its utility or ineffectuality can neither add to nor subtract from this worth. [Kant, p.196 (4:394)]

For Kant, the value of the good will is clearly intrinsic and overriding. But is it universal, objective, and non-proprietary in the appropriate fashion, and how does it matter for establishing his ethical theory? It might seem strange even to discuss what is valuable to humans in the context of Kant's ethical theory, since what humans value is so dependent on the inclinations – wants, desires, feelings, etc. – which Kant dismisses as morally irrelevant. However, such a suspicion elides the distinction between what humans subjectively value, a matter of our inclinations, and what is objectively valuable for humans without regard to our inclinations. For Kant, a good will is manifested when we act out of duty rather than inclination, and only acts performed from duty have genuine moral worth. [Kant, p.198 (4:397)] If individual inclinations have no bearing on whether or not an act is good, then it seems clear that Kant treats the value of the good will as objective in the sense I have used it here.

So the good will is intrinsically valuable, the value of the good will always overrides any other value springing from our inclinations, and the value of the good will is objective rather than subjective. But is it universal in the relevant sense, and therefore non-proprietary? Kant presents multiple formulations of the categorical imperative which he argues are all equivalent to one another, but one of these is the most illustrative for my purpose here:

> If then there is to be a supreme practical principle and a categorical imperative for the human will, it must be such that it forms an objective principle of the will from the idea of something which is necessarily an end for everyone because *it is an end in itself*, a principle that can therefore serve as a universal practical law. The ground of this principle is: *Rational nature exists as an end in itself.* This is the way in which a human being necessarily conceives of his own existence, and it is therefore a *subjective* principle of human actions. But it is also the way in which every other rational being conceives his existence, on the same rational ground which holds also for me; hence it is at the same time an *objective* principle from which, since it is a supreme practical ground, it must be possible to derive all laws of the will. The practical imperative will therefore be the following: *Act in such a way that you treat humanity, whether in your own person or in any other person, always at the same time as an end, never merely as a means.* [Kant, p.229 (4:428)]

In this passage, Kant's usage of 'subjective' and 'objective' is nearer to how I have been using the closely related ideas of 'proprietary' and 'non-proprietary.' To say that rational nature is an end in itself is to say that it is valued for its own sake, not merely as a means to achieve something else of value – that is, our rational nature is intrinsically valuable. In naming the intrinsic value of our rational nature a subjective principle, Kant is saying that every human as a matter of conceptual necessity has a motivating reason to value his or her own rational nature as an end in itself. In naming it an objective principle, Kant is saying that humans must as a matter of *a priori* logical necessity recognize the same value in anyone else's rational nature as in their own, because recognizing rational nature as an end in itself simply means that its value does not depend on any other (subjective) end. In other words, to recognize something as an end in itself is to recognize it as *good* in a non-proprietary sense, not 'good for me' or 'good only for achieving some other end of mine.'

If rational nature is valuable as an end in itself and if any being with a rational nature must recognize it as such (insofar as they are rational), then that rational nature is of intrinsic, objective, universal, non-proprietary value. But what, exactly, is this rational nature that is an end in itself? A few paragraphs before the quotation cited above, Kant wrote:

> We think of the will as a power of determining oneself to act *in conformity with the idea of certain laws*. And such a power can be found only in rational beings. Now, what serves the will as the objective ground of its self-determining is an *end*; and this end, if it is given by reason alone, must be equally valid for all rational beings. On the other hand, something that contains merely the ground of the possibility of an action, where the result of that action is the end, is called a *means*. The subjective ground of desiring is a *driving-spring*; the objective ground of willing is a *motivating reason*. Hence the difference between subjective ends, which depend on driving-springs, and objective ends, which depend on motivating reasons that are valid for every rational being. [Kant, p.228 (4:427)]

The rational will then is the power to determine oneself to act in conformity with the idea of a law which is objective and therefore valid for all rational beings. To say that this power is an end in itself is to say that it is good – not good for a given individual, or good for some other purpose, but simply good – to exercise this power. How does one exercise this power? By acting in conformity with the idea of a law which is objective and therefore valid for all rational beings. A rule for action which applied to some rational beings (me, for example) and did not apply to others would not conform with the idea of any law, since laws are universally valid rather than valid for some and invalid for others: The implied necessity for avoiding self-contradiction by willing two incompatible rules for action is precisely what the 'rational' part of a 'rational will' consists in, for it is necessarily irrational to will at the same time *A* and *not-A*.[18]

Of course, humans often do engage in exceptionalism, setting one standard for their own actions and a very different standard for the actions of others. Kant does not deny this, nor does he deny that our often selfish inclinations motivate our actions. Rather, he argues that we are also capable of acting from reason, and he argues that only actions motivated by and consistent with reason are moral actions – hence the absolute value of the rational will.

What remains, then, is to show that the intrinsic, objective, universal, non-proprietary value of the rational will plays the foundational role of a fact-value bridging claim as I have characterized it. However, it seems difficult to isolate any element of Kant's argument as playing a foundational role because of the interrelated, analytic nature of his argument: He starts with the unconditional value of the good will and ends with it, adding little in between except his particular conceptions of what it means to be a rational being, what it means to will an action, and so on. On the other hand, consider what follows if one rejects any part of Kant's understanding of human nature: What if one is simply not convinced by Kant's account of the character and role of reason in human existence? What if one believed that Kant's account of what "willing" means is flawed, or that his strict separation of reason and inclination is untenable? If one instead believed, with Hume, that reason is and ought to be the slave of the passions (Kant's 'inclinations'), the entire justification Kant offers for his ethical theory would collapse. Kant's understanding of human nature is simply identical with his account of what is

[18] In this limited way, Mill and Kant agree, insofar as their respective ethical principles are the product of reason rather than natural inclination (see p.41.), and insofar as the rational component of each ethical theory is the part that forbids selfishness and demands at least some degree of altruism (by insisting that other agents must be recognized as having moral importance equal to one's own).

universally, objectively, intrinsically valuable to us and why, so to reject his understanding of human nature would necessarily remove the foundation of the arguments for his ethical theory.

Kant, unsurprisingly, is quite aware of this. It is important not to mischaracterize or misunderstand his project, which is a transcendental argument which does not examine morality directly, but rather seeks to establish the conditions for the possibility of morality[19]. Throughout the *Groundwork*, Kant states in several different ways that it is not possible to empirically determine anything about morality.[20] However, this rejection of the empirical realm of facts does not undermine my general thesis about fact-value bridging claims, for it frames Kant's argument for his ethical theory as a very large conditional claim: *If* it is true that humans are rational beings as characterized (which Kant claims we can never empirically discern), *then* morality consists in acting according to the categorical imperative. Since Kant's characterization of humans as rational beings – with a very specific definition of what it means to be a rational being – is essentially identical with his account of what is of objective, intrinsic, non-proprietary value to humans universally, it can and does play the foundational role I've characterized.

Section 5: Virtue theory via Aristotle

In *Nicomachean Ethics*, Aristotle begins his inquiry in much the same place I began my attempt to devise a solution for the fact-value problem (see p.18), with a very broad examination of what value is and what humans value:

[19] Some components of morality are indeed taken by Kant as a given in the sense I referred to in my discussion of transcendental arguments above (see p.31), although he does not establish this clearly in the *Groundwork*. That project is left to Kant's *Critique of Practical Reason*, and is beyond the scope of what I need to discuss about his ethical theory here.

[20] The impossibility of providing any sort of empirical foundation or evidence for morality, not even evidence for a single action motivated solely by duty rather than by the inclinations, hinges on the same distinction that Kant makes between objective 'motivating reasons' and subjective 'driving-springs' in the selection cited above. I will return to this particular issue in Chapter 4 (see p.100).

> Every art and every inquiry, and likewise every action and choice, seems to aim at some good, and hence it has been beautifully said that the good is that at which all things aim. But a certain difference is apparent among ends, since some are ways of being at work, while others are certain kinds of works produced, over and above any being-at-work. And in those cases in which there are ends of any kind beyond the actions, the works produced are by nature better things than the activities. And since there are many actions and arts and kinds of knowledge, the ends also turn out to be many: of medical knowledge the end is health, of shipbuilding skill it is a boat, of strategic art it is victory, of household management it is wealth...
>
> If, then, there is some end of the things we do that we want on account of itself, and the rest on account of this one, and we do not choose everything on account of something else (for in that way the choices would go beyond all bounds, so that desire would be empty and pointless), it is clear that this would be the good, and in fact the highest good. [Aristotle, *NE*, p.1 (1094a, 1-21)]

Effectively, the rest of *Nicomachean Ethics* is Aristotle's account of what the highest good is and how best to achieve it. Even in this first step of framing the question, Aristotle's account incorporates some of the key elements of the solution to the fact-value problem I have proposed. Aristotle focuses on our ends, i.e. what is valuable to humans generally, and he draws the distinction between what is valuable for its own sake and what is only valuable for achieving some other end, i.e. extrinsic vs. intrinsic value. So what is "the highest good" for humans, that which is valued for its own sake and never for the sake of anything else? Aristotle's answer parallels the later utilitarian answer, 'happiness' – but Aristotle gives a much richer and more detailed account of what human happiness consists in than pleasure and the absence of pain.

Going into great detail about Aristotle's account of human happiness and the arguments he makes in developing it would be a distraction from my aim here, however. My only concern is whether and how Aristotle's ethical theory instantiates my proposed general solution to the fact-value problem. Having established from the very start that Aristotle sees human happiness as intrinsically valuable, it remains for me to show that it is valuable in an objective, universal, non-proprietary fashion. That cannot be done without at least some commentary on the substance of his ethical theory, but I will avoid those details of his argument which are not directly relevant to addressing the matters at hand.

Aristotle's ethical theory looks fundamentally different from utilitarian or deontological ethics in that the conclusion of Aristotle's argument is not any sort of principle or law, but an account of what constitutes moral excellence in character. But even if Aristotle presents no straightforward, unary ethical prescriptive statement like the principle of utility or the categorical imperative, he could not give an account of excellence without some standards for judgment: Looking at how Aristotle judges character should reveal whether his standards for judgment have the sort of objectivity and universality required to bridge the fact-value gap. To see this, it is necessary to follow Aristotle's arguments for the central importance of happiness a little bit further:

> But perhaps to say that the highest good is happiness is obviously something undisputed, while it still begs to be said in a more clear and distinct way what happiness is. Now this might come about readily if one were to grasp the work of a human being. For just as with a flute player or sculptor or any artisan, and generally with those to whom some work or action belongs, the good and the doing it well seem to be in the work, so too it would seem to be the case with a human being, if indeed there is some work that belongs to one. But is there some sort of work for a carpenter or a leather worker, while for a human being there is none? Is a human being by nature idle? Or, just as for an eye or a hand or foot or generally for each of the parts, there seems to be some sort of work, ought one also to set down some work beyond all these for the human being? But then what in the world would this be? [Aristotle, *NE*, p.10 (1097b, 20-31)]

Before returning to Aristotle's answer to my previous question, I wish to consider what he has said so far. By analogy both with skilled human activities (flute playing, sculpting, carpentry, etc.) and with parts of the human body (eye, hand, foot), Aristotle suggests a fundamentally *functional* account of human nature. Human life has a *telos* – an end, that towards which something aims, the achievement of which brings completion or wholeness to its activity: When talking generally about ends, Aristotle often uses the analogy of the target for an archer.

[Aristotle, *NE*, p.201-212] Here, Aristotle's analogies about the work of a human imply two different sorts of function – different ways in which an end/*telos* is achieved: These analogues suggest that human beings are constructed by nature so as to fulfill this *telos*, like body parts and their respective functional roles, but also that human beings can become more skilled at achieving this *telos*, like artisans and their respective arts. That said, I will return to Aristotle where I left off:

> ... But then what in the world would this [work for a human being] be? For living seems to be something shared in even by plants, but something peculiarly human is being sought. Therefore, one must divide off the sort of life that consists in nutrition and growth. Following this would be some sort of life that consists in perceiving, but this seems to be shared in by a horse and a cow and by every animal. So what remains is some sort of life that puts into action that in us that has articulate speech; of this capacity, one aspect is what is able to be persuaded by reason, while the other is what has reason and thinks things through. And since this is still meant in two ways, one must set it down as a life in a state of being-at-work, since this seems to be the more governing meaning. [Aristotle, *NE*, p.11 (1097b-1098a, 32-8)]

Here, Aristotle alludes to but does not explain his theory of the soul. Aristotle's understanding of 'soul' (*psyche*) is completely at odds with most contemporary ideas commonly associated with the word. Aristotle defined the soul as the internal principle of motion and rest in anything that moves of itself, arising from and consisting in the arrangement and activity of the parts of the thing – for example, the organs and limbs of an organism. Rather than being the sort of immaterial, immortal, ghostly essence brought to mind by the word today, Aristotle's 'soul' is thoroughly embodied and ceases to exist when the arrangement and activity of the parts is disrupted – for example, when an organism dies.[21]

[21] Concepts change, but linguistic remnants of this older conception of soul are easily found: For example, *anima*, the Latin translation of the Greek *psyche* – a metaphor based on the respective words for 'breath' in each language – is the conceptual and linguistic root of the noun 'animal' and the adjective 'animate' - as opposed to 'inanimate.' Of course, breath is also something that ceases at the end of an organism's life. Further discussion of Arisotle's conception of the soul follows in Chapter 3, Section 3.

With this general understanding of soul in mind, it is easier to understand Aristotle's tripartite conception of the human soul in particular. Humans have an internal principle of motion and rest responsive to nutrition and responsible for growth, the 'vegetative soul' that we share with all other organisms, even plants. Humans also have an internal principle of motion and rest responsive to perception and responsible for movement, the 'animal soul' that we share with animals, but not plants. Finally, humans have an internal principle of motion and rest responsive to reason and responsible for reasoning, the 'rational soul' which we share with other humans, but no other organisms. The rational soul is the part of human nature "that has articulate speech" cited above – that which characterizes us as human and distinguishes us from other animals. Aristotle is saying that the work for a human being cannot be merely living, or living in any manner which we share with many other creatures, but must be something particular to being human: The work of being human is not just living, but living a life governed by reason. So, returning to Aristotle where we left off:

> And if the work of a human being is a being-at-work of the soul in accordance with reason, or not without reason, while we say that the work of a certain sort of person is the same in kind as that of a serious person of that sort, as in the case of a harpist and a serious harpist, and this is simply because in all cases the superiority in excellence is attached to the work, since the work of the harpist is to play the harp and the work of a serious harpist is to play the harp well – if this is so and we set down that the work of a human being is a certain sort of life, while this life consists of a being-at-work of the soul and actions that go along with reason, and it belongs to a man of serious stature to do these things well and beautifully, while each thing is accomplished well as a result of the virtue appropriate to it – if this is so, the human good comes to be disclosed as a being-at-work of the soul in accordance with virtue, and if the virtues are more than one, in accordance with the best and most complete virtue. But also, this must be in a complete life, for one swallow does not make a Spring, nor one day, and in the same way one day or a short time does not make a person blessed and happy. [Aristotle, *NE*, p.11 (1098a, 8-19)]

On this view, the highest good – happiness, the target at which all our actions should aim, the human *telos* – consists in a life not merely lived in accordance with reason in some perfunctory sense, but a life guided excellently by reason. The words 'excellence' and 'virtue' are interchangeable here, both translations for the Greek *aretê* and its derivatives: In general it indicates the quality or qualities which make something an outstanding example of its kind, well-suited to fulfilling its ends, not just functional but functioning very well. The virtues specific to the work of a human being are then the virtues of reason, both the virtues of reason in itself (the virtues of intellect) and the virtues of reason in governing that part of us amenable to reason, i.e. managing our impulses and passions (the virtues of character).

General excellence is not found in single actions, but in a pattern of consistently excellent action. On this point another bit of Greek etymology is very revealing: The words *ethos* and *êthos* can be translated, respectively, as 'habit' and 'character.' Aristotle believed, very plausibly, that humans develop stable patterns of behaving well or badly by repetition of actions: In moral behavior as in so many other things, we learn by doing. Doing what, exactly? Aristotle argued that our various basic motivations and impulses – hunger, fear, lust, and so on – move us towards our own good. Everyone must eat and drink; everyone must exercise caution, and anyone might at some time face danger in defense of oneself and what one holds dear; and there would be no humans if we weren't moved to mate and reproduce. But if we are moved too much or too little by such impulses, our well-being is undermined. Thus, what humans must do excellently to live well and happily is govern those impulses, not allowing ourselves to be moved too much or too little by them – where "too much" and "too little" are judgments made by reason in response to the circumstances of the action, always guided by the *telos* of happiness.

On Aristotle's view, each act of proportional restraint – or lack of restraint, or excessive restraint – in response to some basic impulse contributes to the establishment of the habits that comprise our character, for good or ill. The virtues are those habits – Aristotle called them "stable, active conditions of the soul," but we might call them consistent patterns of inclination and behavior – which lead a person to respond proportionally/rationally to the events in life that prompt reactions in us, neither feeling nor being moved too much or too little by our emotions and impulses. The vices, then, are habits which lead us to respond either excessively or deficiently, feeling and being moved too much or too little by our emotions and impulses.

To see this a bit more clearly, let's take a look at Aristotle's paradigmatic virtue, courage: A person who has developed the virtue of courage does not simply scoff at danger, but rather feels fear in proportion to the danger at hand, and allows this fear to motivate his or her actions only to the degree appropriate to what is at risk from the danger and what is to be gained by facing the danger. A person develops courage by acting courageously – by reflecting and weighing his or her response to individual circumstances of danger, and acting in a manner consistent with courage (moved neither excessively nor deficiently by fear). Habituated by these actions which are guided by reflection, eventually a person acts courageously without the need for reflection, reacting to danger by feeling only as much fear as is warranted and acting accordingly. Those who consistently fail to weigh their response to dangers, and instead allow fear to govern their actions entirely, develop through habituation the vice of cowardice instead of the virtue of courage. Those who instead consistently ignore fear and act without reflection – or who reflect poorly, thus take impulsive action to demonstrate their fearlessness instead of taking a more cautious action proportionate to the danger at hand – develop through habituation the vice of rashness (or foolhardiness) instead of the virtue of courage.

Having said enough about Aristotle's ethical theory to set the stage, I will not enumerate the other virtues or elaborate on them any further. The question at hand is whether and how

Aristotle's conception of the highest good – that which is of intrinsic value to every human – is objective, universal, and non-proprietary.

While the appropriate mean action may differ from person to person and circumstance to circumstance, finding the mean is always the most valuable thing to do: Actions may not always be unambiguously right or wrong on Aristotle's view, but there are definitely better and worse actions, and what makes a given action better or worse is a matter of objective fact rather than subjective preference. Moreover, it is always true that using reason to find the mean is valuable, and letting one's passions drive one to excess or deficiency is disvaluable.

This view constitutes more than mere *pro forma* objectivity; it has content. On Aristotle's view, better actions – actions where one uses reason to find the mean between being excessively or deficiently moved by ones impulses (desires, passions) – do as a matter of fact lead to better character (virtues) through habituation, and worse actions do as a matter of fact lead to worse character (vices). Such a conception of the good may not be neat or precise, but it is nevertheless objective in the required sense: On Aristotle's view, someone who believes that self-preservation is more important than anything else and always avoids or flees danger is simply mistaken about what is most valuable, as is someone who laughs at every danger and seeks out unnecessary and completely avoidable risks to life and limb for the thrill of it.

Aristotle also allows play for individual and circumstantial variance without sacrificing a broad universal basis for value. While humans may differ substantially in our individual capacities and talents, the capacity for reasoning – both the ability to engage in reasoned reflection about our actions and the ability to govern the impulses that motivate us to act – is a universal human characteristic, the very thing that sets us apart from other organisms. Similarly, the value of a reason-governed life is universal: All humans have the capacity to exercise their reason to govern their impulses, and it is valuable for every human to do so.[22]

Excellent character is universal in the sense that the being-at-work of a human being is the same for all human beings, but that is not quite equivalent to one and the same thing being valuable to all humans: It leaves open the possibility that what is valuable to every human is the excellence of his or her own character – not virtue in general, but one's own virtue. This leaves open the possibility of a proprietary perspective on the value of a reason-governed life: That is, it is clearly of value to me that I should exercise my reason to govern my impulses, that I should seek the mean and develop the virtues. But is it of value to me that other people should be virtuous?

First, it should be noted that phrasing the question in such a way is a bit misleading: In fact, other people's virtue being of value specifically "to me" would be just another way for virtue to be proprietary. It may not be in my personal interest that other people be virtuous, but my recognition of the value of a life governed by reason does not depend in any way on it being my life in particular. On that basis, the argument explaining the non-proprietary value of happiness for Mill (see p.40) would also seem applicable to Aristotle's ethical theory: Even if I don't have a vested interest in the happiness of others, reason may require me to recognize that when someone else acquires the virtues and leads a human life well-lived, he or she has accomplished something of intrinsic value. From this perspective, as with Mill and Kant, it is reason rather than natural inclination which demands that others be recognized as having moral importance equal to one's own (see p.41 and footnote on p.49).

[22] More honestly, Aristotle believed that every *male* human had these capacities. For all his independent-mindedness in other matters, there is no evidence that Aristotle ever eschewed or transcended ancient Greek cultural attitudes towards women.

Aristotle, however, does not make this argument: Moreover, unlike Mill's case for utilitarianism, the argument Aristotle *does* make seems not to imply or require this particular maneuver. Mill's argument hinges on perceiving happiness as desirable/valuable, and the recognition that happiness has non-proprietary value was motivated by the question of exactly what is perceived – that happiness is valuable, or that happiness is valuable-for-me, the latter of which would defy reason. There is no aspect of Aristotle's argument that seems to hinge on a similar question of perception and perspective – so the "reason must recognize...they have accomplished something of value" argument may be applicable, but it does not follow naturally from the basis of Aristotle's argument the way it does from Mill's.

So, to ask the question again in a less misleading fashion: Is the virtue of other people of intrinsic value to any given individual within Aristotle's virtue theory? There are a couple of different lines of argument which suggest this Aristotle's highest good does include a component of universal/non-proprietary value. Firstly, Aristotle maintains from the very beginning of his argument that inquiry into the nature of the highest good for man properly falls under the art of politics – which rather strongly implies that the highest good is collective. That is, if the highest good is a political matter, my highest good would include not only my virtue but my fellow citizens' virtue. Secondly, Aristotle's development of the virtues of justice and friendship both focus on practicing the other virtues (courage, generosity, etc.) towards other people – again implying a fundamentally collective understanding of the highest good.

However, Aristotle's conception of politics is not itself universal – it is constrained by his conception of the *polis*, which Joe Sachs describes as "a self-sufficient political community, large enough to feed and defend all its members but small enough for them all to have active dealings with one another." [Aristotle, *NE*, p.2] Similarly, justice is practiced with respect not to humans in general, but with respect to one's fellow citizens – and friendship is presumably practiced with respect to a smaller subset of humanity than that. So it is not entirely clear that Aristotle's conception of the highest good for humans is in fact non-proprietary in the fullest sense.

I do not see this ambiguity about whether or not Aristotle's conception of the highest good is fully non-proprietary as a complete failure to establish that Aristotle's ethical theory instantiates my proposed general solution to the fact-value problem. At least in part, it stands as an open question about Aristotle's success in advancing a truly universal ethical theory: Within the scope of ordinary human life in a community with others, Aristotle's ethical theory is non-proprietary. If it is not entirely clear whether the prescriptive claims of Aristotle's ethical theory specify right behavior with respect to all persons, i.e. universally, then it is equally unclear whether or not Aristotle's ethical theory actually satisfies the minimal standard for an ethical theory I described as the working assumption for my overall conditional argument (see p.31+).

In other words, if closer analysis of Aristotle's arguments were to reveal that his conception of the highest good is truly non-proprietary, then his ethical theory would instantiate my solution to the fact-value problem. And if such an analysis were to reveal that Aristotle's conception of the highest good is less than fully non-proprietary, it would also to that extent raise questions about whether it even counts as a legitimate ethical theory: That is, if proprietary, Aristotle's ethical theory's answers to the question "How should one live one's life?" would only be substantial with respect to one's treatment of one's fellow citizens, and not address conduct towards other humans at all – meaning that the only moral rule with respect to the bulk of humanity would be, in effect, "Do as thou wilt."

Since either of these conclusions are satisfactory with respect to my overall argument, I don't want to spend time arguing for one or the other here. And, as will become evident over the course of the next chapter, the extent to which an individual's *telos*/highest good encompasses the highest good of others in a non-proprietary fashion will be a productive area of inquiry in the development of my own attempt to ground a conception of the highest good in an evolutionary understanding of human nature.

In summary, Aristotle characterizes humans as beings whose highest good – the target at which all our actions should aim, our *telos* – is happiness, which consists in a life guided excellently by reason. Aristotle defines what constitutes excellent guidance by reason in multiple ways which may themselves be problematic, insofar as they may rely on further norms or value claims which are not independently justified. However, such a critique lies beyond the scope of my interest here. For my purposes, what matters is that the *telos* of happiness so defined is Aristotle's fact-value bridging normative claim, and is perhaps the clearest example of one so far: To name some state of affairs a *telos* is, as Aristotle defines the term, simply identical with claiming that achievement of this state of affairs is of objective, intrinsic value to every being for whom it is the *telos*. As such, a *telos* does not merely imply or define value, it determines what else is valuable to that which has the *telos*.

However, it is worth noting that there is nothing in the definition of *telos* that directly requires or implies that the values determined by a *telos* must be non-proprietary: That requires additional argument, even within Aristotle's own ethical theory – an argument that is unresolved. A being's *telos* is its highest good, and as such is the pinnacle of a nested hierarchy of goods encompassing everything that is good for/valuable to that being: But whether that *telos* establishes a basis for genuinely non-proprietary value claims may depend a great deal on the specific nature of the *telos* in question.

CHAPTER 3

EVOLUTION, TELEOLOGY & VALUE

Section 1: Re-framing the problem of value

Thus far, I have established the criteria which any fact-value bridging normative claim must meet in order to overcome the fact-value problem and provide a firm foundation for ethical theory. Subsequently, I endeavored to show how the three of the most influential ethical theories in the history of philosophy do in fact rest on fact-value bridging normative claims which satisfy those criteria. In each case, those fact-value bridging normative claims were themselves justified within an account of human nature. Ultimately, then, the ethical theories built on these foundations can be no more convincing than their underlying account of human nature: If we are not in fact creatures for whom happiness is the only thing valued as an end in itself, the foundations of utilitarianism are built on quicksand. Likewise for deontological ethics if Kant's account of the nature of human reason, inclination and will is flawed; and for virtue theory if Aristotle's teleological account of human nature is mistaken. Hence the motivation to see whether and how some fact-value bridging normative claim might be formulated and justified within an account of human nature rooted in the firmer ground of science.

But even if the motivation is clear, the method is not. To build a bridge, it is necessary to find the right point of connection, a narrow place between the realms of ethics and science where the bridge might be built. It isn't difficult to see that there are, and indeed must be, places where the two realms connect: It is hardly possible to address any issue of substance in applied ethics without confronting human biology in the form of our needs and limitations. There would be no prohibitions against torture or murder if we were not mortal creatures who suffer and die, no obligation to feed the hungry if we did not hunger.

In ethical theory as opposed to applied ethics, however, the brute facts of human need rooted in our biology generally take a back seat to some aspect of cognition, broadly conceived: Utilitarianism is concerned with the pleasure we experience in satisfying our needs and desires and the pain suffered when needs and desires go unsatisfied; the needs themselves are but background, a source for some of our pleasures and pains. From a Kantian perspective, our desires are mere inclinations, and even needs are but a source for inclinations – and inclinations are morally irrelevant except insofar as actions springing from them are also the objects of our will: Kantian morality consists solely in willing rationally, which is completely abstracted from any of the particular needs or inclinations which motivate the actions so willed.

Aristotle, in contrast, took our biologically rooted needs and desires very seriously, and viewed even cognition and experience – including our capacity for reason – in biological and functional terms. From an Aristotelian perspective, our impulses and inclinations are valuable to us in a very straightforward way: For example, we experience the sensation of hunger because awareness of our bodily needs is required for us to sustain our lives. We have the capacity to take pleasure in food because pleasure motivates us to satisfy those needs and sustain ourselves. Awareness and motivation are animal capacities as well, and Aristotle makes no distinction between humans and other animals in these capacities. Instead, the distinction Aristotle draws between humans and animals rests entirely in our capacity to reason, which animals lack. We can weigh our options and choose what is best for us, which includes choosing how to respond to our animal awareness and motivation so that we do not, for example, eat too much or too little for our health and well-being.

In the search for a fact-value bridging foundational claim, the most important lesson to be drawn from Aristotle lies in his approach to building that bridge: From an Aristotelian perspective, there is no bridge to build because the facts of human nature are already laden with value. Indeed, none of the ancient Greek philosophers recognized the fact-value dichotomy that structures so much of modern ethical theory focuses on, and which I have worked here to overcome. On Aristotle's view, we are by nature beings with ends, and examination of our various ends reveals a nested hierarchy of ends which culminates in an ultimate end, our highest good, our *telos*. And not just ours, for other beings also have their own *teloi* which determine what is of value to them: In *Nichomachean Ethics*, Aristotle is concerned with determining the highest good for a human, but his argument implies that there is a highest good for a horse or a hummingbird as well, albeit not a highest good involving reason.

My strategy will be to learn from Aristotle, and to develop an understanding of *telos* grounded in evolutionary biology rather than Aristotelian biology. This is dangerous territory, however: For all that Aristotle was history's first systematic biologist and has been rightly called "the father of biology," the relationship between modern biology and Aristotelian biology is strained at best. To most modern biologists and philosophers of biology, Aristotle is less a father than an elderly uncle – one holding outdated and somewhat embarrassing opinions which the younger generation would just as soon ignore. I will address why that is the case in the next section: However, my primary focus will not be to explain this division. Rather, my intent is to overcome it.

Section 2: Teleology and the essence of life

> Now such a minimal list of such maximal centrality and importance bears a description in ordinary language – but its proper designation requires that evolutionary biologists utter a word rigorously expunged from our professional consciousness since day one of our preparatory course work: the concept that dare not speak its name – essence, essence, essence (say the word a few times out loud until the fear evaporates and the laughter recedes). It's high time that we repressed our aversion to this good and honorable word.
>
> – Stephen Jay Gould, *The Structure of Evolutionary Theory* [p.10]

The reason for the expurgation of the word 'essence' from biology, alluded to by Gould above, has to do with the conceptual history of biology and the legacy of Aristotle. The Aristotelian view was that a species or kind is defined by its essential characteristics, the distinctive features shared by all the individual members of the species. Differences between individuals have little or no place in such a view – by definition they are inessential characteristics, at most unimportant variations in the expression of essential characteristics. It is only shared characteristics which make each kind what it is, and organisms reproduce according to their kind.

For the process of evolution by natural selection to be recognized and understood, this flawed species concept had to be overcome: Until one recognizes that a species consists in a population of genuinely differing individuals, one cannot possibly see how those differences between individuals can result in differential reproductive success, and how the inheritance of those differences over generations of differential reproductive success can change the characteristics of whole populations – that is, evolution by natural selection. This is why Darwin spent the first two chapters of *On the Origin of Species* discussing the evident variation between individual organisms within a species. [Darwin, p.7-59]

So before they could understand natural selection, 19th Century natural historians had to abandon essentialism and learn to see species as populations of genuine individuals with differences that truly *matter*. But what differences matter, and why do they matter? Those differences which cause individuals to differ in their reproductive success. And herein lies an essence – not the essence of species, nor even the essence of individual organisms, but the essence of life itself.

One way to get at the essence of life is to focus on what constitutes an essence at all. The concept of an essence rests on the idea of an essential property, a property or properties that something simply must have – or at least properties that it must have in order to be the kind of thing that it is. This is opposed to accidental properties, those properties that something happens to have but need not have – or properties that are not defining features which characterize its kind. Some philosophers argue that there are no essential properties, and others argue for a different (non-modal) definition of essential and accidental properties, and so on – but none of those disputes are germane to the matter at hand. Slightly more relevant is modern biology's already-noted position that species have no essential characteristics, that it is impossible to specify any set of characteristics which would unambiguously determine which individuals are members of a given species and which are not. I do not dispute that claim in the least: Species have no essence in this sense, although some species may happen to have – "accidentally," as it were – distinctive characteristics shared with no other known species, such as the extraordinary necks of giraffes, or the capacity for language in humans. The definition of what constitutes a species is fluid because the phenomena the definition is intended to describe are complex and varied, as are the criteria for membership within a given species – especially over spans of time and space.

But agreement that species do not have essences in this sense does not require or imply the claim that there are no essences, or that life itself has no essence. So is there a common characteristic that is essential to every living thing, a property a thing must have in order to be a living thing, a property which unambiguously determines which things are living things and which are not? I believe so, and here is a deceptively simple way to articulate it: **The one property every organism shares with every other organism is that all of its ancestors reproduced.**

While that may seem like an empty tautology – it is surely implicit in the meaning of the word "ancestors" – it is not such an empty claim when one considers all the organisms that are not the ancestors of any living thing. At every moment in time from the earliest beginnings of life on earth to the present day, many organisms have existed which did not reproduce and so did not become ancestors to any of the organisms in succeeding generations: If we pare the notion of a lineage of organisms down to its core – and the lack of any essential properties which determine species membership provides at least some motivation for doing so – then every organism that dies without reproducing itself is an extinction event of sorts. Some (perhaps even an overwhelming majority) of those failures to reproduce are the result of simple bad luck, but many of them represent natural selection in action – and every organism that exists shares equally in the distinction that its lineage has not yet gone extinct, that it is the product of billions of years of successful reproduction. Surely that puts reproduction and natural selection right at the heart of understanding the essence of life.

The "minimal list of such maximal centrality and importance" that Gould refers to above is a list of the central defining characteristics of the theory of evolution by natural selection:

> The basic formulation, or bare-bones mechanics, of natural selection is a disarmingly simple argument, based on three undeniable facts (overproduction of offspring, variation, and heritability) and one syllogistic inference (natural selection, or the claim that organisms enjoying differential reproductive success will, on average, be those variants that are fortuitously better adapted to changing

> local environments, and that these variants will then pass their favored traits to offspring by inheritance). [Gould, p.13]

Gould goes on to talk much more about evolutionary theory and its conceptual core – the efficacy of the process of natural selection to produce the organic diversity we see, the scope of the theory to account for the bulk of that diversity, the fact that organisms themselves are the agents of evolutionary change – all of which will be reflected in my discussion to follow. But for my purposes here, I am not as interested in the essential core of the theory of evolution as the essential core of that which evolves, the essence of life itself.

In order to isolate and articulate the essence of life implied by the facts of evolution, it might be useful to consider how these bare bones of natural selection generate equally bare-boned results. To do this, I will set aside the concrete features of actual organisms and examine the nature of an abstract entity with the minimal set of properties consistent with the bare bones of natural selection – a thought experiment designed to tease out the implications of the essence of evolution for the essence of life.

The "three undeniable facts" that Darwin recognized and Gould cites tell us a great deal about the properties that any entity must have in order for natural selection to operate on it: If an entity produces offspring at all, no matter how loosely we interpret the concept of offspring, it must at minimum have the capacity to produce other entities. If different individual entities vary from one another, they must have different properties – either other properties beyond the minimally required capacity to produce other entities, or differences in how that capacity operates – but we can leave these different properties otherwise unspecified. If the entities exhibit heritability, an entity's properties – both its capacity to produce other entities and any other unspecified properties it has – must also be exhibited by its offspring entities: This requires that an offspring entity must not just be a production of its parent entity, but a reproduction – a copy of the original.

But note that I started by hypothesizing an individual entity until I moved on to the matter of offspring and variation. If our abstract entity engaged in perfect-fidelity reproduction every time, how could any variation in the properties of a population of entities ever have arisen in the first place? Combining the fact of variation with the fact of heritability implies at least some imperfection in reproduction. To satisfy all of the "undeniable facts" at the heart of evolutionary theory, our hypothetical entity must reproduce itself with at least an occasional addition, subtraction, or alteration of some property or properties, which differences come to constitute the varieties of subsequent generations of entities. Because nothing about this reasoning requires or implies any particular constraints on these additions, subtractions, or alterations of properties, it seems best to think of them as random.[23]

For these abstract (imperfectly) reproducing entities to not just produce offspring but *overproduce* them, one must suppose that there are constraints on reproduction; for example, there might be resources required for reproduction that are scarce in the local environs. It is the constraints on the system of reproducing entities that constitute criteria for the filtering process of selection: If there are constraints on a population of entities with varying properties which prevent them all from reproducing themselves indefinitely, then those entities which vary in some way that gives them an advantage in reproducing themselves (within those constraints) will eventually outnumber those variants at a comparative disadvantage.

[23] It seems worth noting here that Darwin conceived the theory of natural selection knowing little about the actual mechanisms of reproduction. In this abstracted account of nature, I am deliberately assuming even less knowledge than he had.

Thus, even when considering purely abstract and entirely hypothetical entities whose only specified properties are those necessary to exhibit the three undeniable facts about living things specified in the Gould quotation cited above, the syllogistic inference of evolution by natural selection follows. Even if we were to imagine infinite external resources and space for our hypothetical entities to reproduce themselves, a constraint so slight and internal as each entity taking a certain finite amount of time to reproduce itself or to rest between reproductive cycles could be a source of variation subject to selection: Under such a minimal constraint, the entities which take slightly less time to reproduce each generation would eventually come to vastly outnumber those which take slightly longer, at least over the course of many generations.

Nowhere in this analysis of the basic apparatus of natural selection did I limit the number, type, or nature of the varying properties our abstract entities might have, except that they be heritable (albeit not perfectly heritable). However, the process of natural selection itself limits and alters the character of the properties exhibited by these self-reproducing entities over time: Any properties which positively contribute to successful reproduction become more prevalent in future generations, and any properties which negatively impact reproduction become less prevalent – even if we interpret "positive" and "negative" impacts only in the sense of comparing the entities with each other, as in the example of shorter and longer reproductive cycles.

As new properties arise through whatever imperfections exist in the workings of inheritance, those too will be filtered by selection if they have any impact whatsoever on reproductive success. Given many generations, properties which cause even slight positive or negative comparisons of reproductive success between individuals should spread or diminish throughout the population, especially properties with an impact on reproductive success across changing constraints.

Why changing constraints? If I am to adhere rigorously to the conditions of my thought experiment, introducing no assumptions about my abstract entities besides their adherence to the "three undeniable facts" about living things, then reproductive constraints are subject to change: The only limit on such constraints which follows from the fact of overproduction of offspring (from which the existence of constraints was deduced) is that our entities must be able to produce offspring in order to overproduce them – that is, the constraints cannot be so strict that no entities can reproduce at all. But even if the constraints did become so strict as to prohibit reproduction entirely *on occasion*, as long as those strict reproductive constraints obtained only within a limited location or time span, variants capable of enduring those times intact or escaping to less constrained places would continue to reproduce. This reasoning reveals that constraints which vary across space and time can favor or disfavor quite different variants: There is a sort of feedback between constraints and variants, such that a given circumstance might not constitute a constraint at all for one variant but might entirely prevent the reproduction for another variant.

Changing constraints are not necessitated by the conditions of this thought experiment, but are certainly permitted by them; and they seem almost inevitable if one allows that some constraints on reproduction might be the actions of other entities (resource competition, scavenging, predation). Variable constraints enrich and complicate the system of evolving entities, causing diverse lineages of entities with varying properties to arise rather than encouraging the development of a single dominant entity with properties maximized for reproductive success. Indeed, there can be no such maximization when constraints vary over space and time, and when variants themselves vary with respect to whether and how their reproduction is limited by any given constraint. So even for these minimally specified, purely hypothetical entities, attention to the core facts of life implies the evolution of a rich and complex tapestry which recalls the closing words of *The Origin of Species*: "...from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved." [Darwin, p.490]

But even in an increasingly rich and complex system of diverse lineages of variant entities generated in response to local and ever-changing constraints, those entities' properties will be filtered by natural selection with respect to their contribution to reproductive success under the range of constraints applicable to a given variant. No matter how much the constraints themselves might change, there is but one criterion by which varying properties of our entities are selected – reproductive success. And this is the unifying characteristic that all of the varying properties of these abstract entities/organisms share, the essence of life itself: Over generations spanning changing constraints and including new variations introduced by imperfect heritability, the properties of the entities are increasingly those which are *good for reproduction* – in other words, the properties of the entities become directed towards an end; they become *teleological*.

One can even imagine with no great effort a nested hierarchies of functional properties that such an entity might have: An entity might have a body especially streamlined for moving through its medium, which property is good for mobility, which is good for escaping predators and/or capturing prey, which is good for increasing the entity's odds for survival, which is good for increasing its odds for reproducing itself and passing on that streamlined body to another generation – which is how it got streamlined in the first place, since any ancestors slightly more streamlined than others in its population would have had a reproductive advantage. Since an entity's own reproduction is the level at which properties are filtered by natural selection (at least as we have investigated it so far), reproduction must be the highest good in such a hierarchy: All of which suggests that for any entities which satisfy the requirements under which natural selection can and must occur (variation, heritability, and overproduction of offspring), the operations of natural selection inevitably produce entities with a *telos*. Since every living organism with which we are familiar is in fact the product of billions of years and uncounted billions of generations of selection based on reproductive success, the *telos* of every living organism is its own reproduction.

Nothing in this argument requires that every property that an entity has – or, leaving my abstracted hypothetical example behind, every trait that an organism has – must be the result of selection: A given trait might have no impact whatsoever on reproduction under any constraints a given population of organisms currently faces, and possibly even no impact on any constraint the organisms' ancestors ever faced. I will call such a trait 'selectively neutral.' If a trait is selectively neutral, whether a given member of the population has or lacks that trait will have no bearing on its reproductive success, so the trait might persist in a lineage of organisms indefinitely once it arises – or it might disappear for no particular reason but chance, a victim of genetic drift.

A trait might also be somewhat selectively neutral if it has a (comparatively) negative impact on selection under some constraints and a positive impact under other constraints: If those negative and positive impacts are balanced – as, for example, when the differing constraints are ones that a population regularly faces in more or less equal proportions – the trait might persist in the population in much the same way as a trait with no reproductive consequences at all. Such a trait may be less likely to be eliminated entirely by drift because selection is still operating on the trait: The potential for selection to oppose drift might depend on how regular the to-and-fro tug of positive and negative selective pressure is, so we might be better off referring to such a trait as 'balanced selectively neutral' rather than just selectively neutral.

Selectively neutral traits might arise by purely contingent happenstance, due to a random mutation (such as a gene duplication with no immediate consequences). Significantly, the possibility of selectively neutral traits which arise for no particular reason later becoming subject to selection under different constraints is the engine for evolutionary innovation; such a trait is called an 'exaptation' by Gould and others. But for any given selectively neutral trait, exaptation may or may not ever happen: Some traits might originate and remain selectively neutral indefinitely.

A selectively neutral trait might also arise as a side effect of some other trait that has undergone selection: If causally paired to a trait that does undergo selection, a selectively neutral trait can even become more or less prevalent in the population as if it were itself undergoing selection, yet quite without regard to whether the trait itself has any consequences for reproductive success.

All of these complications – especially the last one – make it extraordinarily difficult to determine whether any given trait is an adaptation or not: For any given trait, it may be there because it helps the organism realize its reproductive *telos*; or it may just happen to be there more or less randomly; or it may have positive and negative reproductive consequences which depend in turn on varying external constraints; or it may be causally connected to a trait that has selective benefits even though it has no particular particular impact on reproduction of itself. But the claim that the *telos* of every organism is its own reproduction does not imply or require that every aspect of an organism – every physical property, every causal capacity, every behavior – is a product of selection and has a functional role. Nor does the claim that something has an essence require that all its properties be essential properties – which would rather miss the point of the distinction between essential and accidental properties.

All that is implied or required by this argument is that *some* of an organism's properties are as they are because they are *good for* reproduction, the organism's *telos*. In reality, I don't think that any working evolutionary biologist would attempt to deny that a great majority of the discernible features of any given organism have in fact been shaped by selection, from the tiniest inner workings of each individual cell through the most complex behaviors of social animals towards one another. But it is not merely the possession of adapted traits that makes an organism an organism: Rather, it is possession of the potential to undergo (and the history of its ancestors having undergone) the process of progressive adaptation by natural selection that characterizes living things – and this potential is inherent in and springs from every organism's drive to reproduce itself, the *telos* of reproduction. This *telos* is the essence of a living thing, what makes it the kind of thing it is.

Section 3: Aristotle and Darwin; *telos* and evolved value

Unfortunately, Aristotle's teleology is no more popular with modern biologists than his essentialism, for reasons I will discuss below. Popular or not, concepts which look very much like teleology to anyone familiar with Aristotle have never been eliminated from biology. What is an adaptation but a trait that is *good for* an organism's reproductive success? While the concept of some property or action being "good for" some discernible end is not all there is to Aristotle's conception of *telos* by a long shot, it is certainly close to the conceptual heart of the matter. On that basis, it is reasonable to suppose that something like an Aristotelian perspective on value is not so unapproachable from the perspective of modern evolutionary biology after all.

But, as I said, teleology is not popular with modern biologists – which has led to various attempts to purge the lingering teleological elements of biology, to rethink and reformulate them so that it can be claimed that they only *appear* teleological, but really aren't. These efforts fail, I think, because they are rooted in deep misunderstandings of what a *telos* is and how teleology operates. For example, the great evolutionary biologist Ernst Mayr explicitly renounced teleology, renaming and reexplaining some of the apparently teleological elements of biology in the following ways:

> Teleonomic activities – "The discovery of the existence of genetic programs has provided a mechanistic explanation of one class of teleological phenomena. A physiological process or a behavior that owes its goal-directedness to the operation of a program can be designated as 'teleonomic'... [teleonomic activities] are characterized by two components: they are guided by a program, and they depend on the existence of some endpoint or goal that is foreseen in the program regulating the behavior." [Mayr, p.48]
>
> Adapted systems – "It was one of the most decisive achievements of Darwin to have shown that the origin and gradual improvements of... organs could be explained through natural selection. It is therefore advisable not to use the term teleological ('end-directed') to designate organs which owe their adaptedness to a past selectionist process." [Mayr, p.50]

Given the conceptual analysis of the core of natural selection that I developed above, it seems clear that there is still a *telos* at work in both of these "non-teleological" elements of biology: In the former, the *telos* is "some endpoint or goal that is foreseen[24] in the program regulating the behavior." In the latter, the *telos* has already been met by the organism's ancestors in "a past selectionist process." In both cases, what is the true *telos*? Reproduction.

Mayr also explains two other teleological elements in scientific thought, which are instructive in a different way:

> Teleomatic processes – A process that reaches a definite end state through the operation of physical laws is teleomatic. For example, "When a falling rock

[24] 'Foreseen,' of course, is not meant literally. It is simply meant to indicate that a step-wise, or 'algorithmic,' causal process (such as a genetic developmental "program") is structured in such a manner so as to proceed to the specified endpoint. The very fact that such ambiguous and misleading phrases seem to naturally find their way into such discussions is perhaps the clearest indicator of why many (including myself) consider biology to be irreducibly teleological.

> reaches its endpoint, the ground, no goal-seeking or intentional or programmed behavior is involved, but simple conformance to the law of gravitation." [p.49]

> Cosmic teleology – "...[T]wo thousand years before the proposal of the theory of natural selection, Aristotle could think of only two alternatives when encountering instances of adaptation: coincidence (chance) or purpose. Since it cannot be coincidence that the grinding molars are always flat and the cutting teeth (incisors) sharp-edged, the difference must be ascribed to purpose. 'There is purpose, then, in what is, and in what happens in Nature.' Indeed, so much in the universe reflects seeming purpose that final causation must be postulated... It is this teleology which modern science rejects without reservation. There is not and never was any program on the basis of which either cosmic or biological evolution has occurred." [p.50]

The lesson from Mayr's description of teleomatic processes and cosmic teleology is two-fold: The first lesson is simply that sometimes eliminating teleology is the right thing to do. As far as physics goes, there is near-universal agreement that Newton's universal gravitation is vastly superior in every way to Aristotle's teleological physics.

The other lesson lies in the notable contrast between the two sets of concepts: In his descriptions of teleonomic activities and adapted systems, Mayr is clearly attempting to provide an alternate explanation for what still appears to be thoroughly teleological even within modern biology. The phenomena are still there, and they are still apparently teleological, and perforce the goal of eliminating teleology requires that Mayr attempt to re-interpret them so they seem non-teleological. In contrast, Mayr forthrightly declares teleomatic processes and cosmic teleology to be false and misleading ideas. Again, the phenomena to be explained can still be seen, but the former teleological explanations for those phenomena have been completely replaced by better explanations lacking any hint of teleology. In light of the better explanations now available, even the phenomena themselves lack the appearance of teleology, so no re-interpretation is required.

My view is that nothing is accomplished by Mayr's (and others') attempts to reconfigure and reinterpret teleology in other terms but to obscure the true *telos* at the heart of any living thing, the drive to reproduce itself. But why such a strong desire to purge *telos* in the first place, especially since the reproductive drive itself is hardly controversial within the framework of evolutionary biology? The need to reject Aristotle's essentialist concept of a species raised in the prior section would seem to be a separate issue, and doesn't necessarily explain the hostility to teleology. The failure of teleology in the physical sciences provides another motive, albeit not a particularly direct one for the biological sciences.

In the end, Darwin's recognition that the superfecundity of nature (overproduction of offspring) generates brutal competition and at best a dynamic equilibrium rather than the more comforting but illusory natural harmony produced by divine purpose seems to be the most plausible primary motive for eliminating teleology from biology: Evolution by natural selection eliminated both the appearance of and the explanatory role filled by cosmic teleology. After that successful elimination of a mistaken *telos*, any attempt to develop a clear understanding of the remaining irreducibly teleological elements of life became the proverbial baby tossed with the dirty bathwater of cosmic teleology and nature's divine harmony.

Even beyond the unwarranted affiliation of all teleological explanation with one particular false teleological explanation (cosmic teleology), much of the rejection of teleology seems to hinge on simple misunderstandings of what *telos* actually means. Mayr provides an example of one common misunderstanding when he feels it important to say that "the existence of genetic programs has provided a mechanistic explanation of one class of teleological phenomena" [Mayr, p.48 as cited above]. This implies that teleological explanation is by its very nature opposed to a mechanistic explanation, based on the presumption that every *telos* is the product of intention or will – as it is presumed to be in cosmic teleology. That is not the case.

Telos has loosely been translated as 'end' or 'goal.' Goals are products of consciousness, inherently intentional, and we are familiar with them because, simply put, we have goals. Thus, speaking teleologically about anything other than ourselves is seen as inappropriately anthropomorphic, imposing a human character on the world that may not really be there (and probably isn't). Moreover, we set goals and change goals all the time, so they seem to have a certain inherent arbitrariness. Therefore it is seen as doubly wrong to talk about goals in the non-human world, because they may not be there at all and because they are inherently arbitrary.

The translation of *telos* as 'final cause' is just as likely to cause concern, if not more so. Neither the modern understanding of the word 'cause' nor ordinary experience in the world allow for a future state of affairs causing anything in the present, except perhaps at the more bizarre and counterintuitive outer limits of theoretical quantum physics.

But translating *telos* as 'goal' or 'final cause' is conceptually loose, and tends to foster inappropriate implications at every turn. In his translation of and study guide to Aristotle's *Physics*, Joe Sachs suggests that 'completion' is a more apt English substitute than either of these [Sachs, p.246]. 'Completion' has none of the anthropomorphic implications of inherent intentionality and arbitrariness that 'goal' has, as will become more clear in the discussion of Aristotle's biology that follows. Similarly, the idea that a step-by-step causal process might arrive at some state of completion – or that an algorithm might be structured so as to arrive at some particular state specified in the algorithm itself – doesn't have the bizarre temporally backward causation implications that might dog the phrase 'final cause.'

Even if some of the rejection of Aristotelian teleology exhibited by modern scientists in general and biologists in particular is based primarily on misunderstandings, or on extensions of valid criticisms beyond their proper scope, it remains for me to give some sort of clear account of how and in what manner Aristotelian biology and evolutionary biology might converge. For that, I must explore Aristotle's understanding of the essence of life in more detail.

Aristotle's account of the nature of living things hinges on his concept of *psyche*, or soul, which is described in greatest depth in his work *On the Soul* (often referred to by its Latin title, *De Anima*). As I had occasion to note in my discussion of *Nicomachean Ethics* above (p.53), Aristotle's concept of the soul bears no resemblance to the modern notion of a mysterious, immortal, supernatural entity that is separable from but attached to a person. Aristotle defined the soul as the internal principle of motion and rest in anything that moves of itself, arising from and consisting in the arrangement and activity of the parts of the thing – for example, the organs and limbs of an organism. As such, *psyche* is simply a property of organisms, inseparable from their physical bodies:

> ... [W]e can dismiss as unnecessary the question whether the soul and the body are one: it is as though we were to ask whether the wax and its shape are one, or generally the matter of a thing and that of which it is the matter. Unity has many senses (as many as 'is' has), but the proper one is that of actuality. [Aristotle, *On the Soul*, p.657 (412b, 5-9)]

So, for Arsitotle, the *psyche* is the actuality of a living thing: But what exactly is "actuality"? The Greek word traditionally translated as "actuality" is *entelecheia*, and this particular translation has caused much confusion over the centuries. Joe Sachs translates the word usefully, if somewhat awkwardly, as "being-at-work-staying-itself." The glossary of his translation of Aristotle's *Physics* is helpful in understanding exactly what this strange hyphenated phrase means:

> being-at-work-staying-itself (*entelecheia*): A fusion of the idea of completeness with that of continuity or persistence. Aristotle invents the word by combining *enteles* (complete, full-grown) with *echiein* (= *hexis*, to be a certain way by the continuing effort of holding on in that condition), while at the same time punning on *endelecheia* (persistence) by inserting *telos* (completion). This is

> a three-ring circus of a word, at the heart of everything in Aristotle's thinking...
> [Sachs, p.245]

While *entelecheia* is an extremely complicated concept, the main thrust of the idea is exactly what makes 'actuality' a poor translation: 'Actuality' has very static implications, and Aristotle's concept of nature is fundamentally active. Things don't just passively exist; things are what they are by virtue of what they actively do – and do for a reason.

What reason? What is the character of these activities of living things? They are all activities that a living thing engages in to keep on being what it is – in short, to survive. The thing does what it does in order to be keep on being what it is, and it is the self-referential character of this *entelecheia* that makes 'actuality' a fair translation in some respects: A living thing is complete, a whole unto itself. What the word 'actuality' misses is the notion that a thing isn't just statically a whole unto itself, but that it must constantly act in order to retain this wholeness.

When stated abstractly, this notion seems very strange to us: A thing is most complete when it constantly acts in order to remain complete? What action does a rock or a chair take to remain what it is? How bizarre to think that a boulder must be "doing something" simply to keep on being a boulder![25] But when these same concepts are applied to living things, they don't seem bizarre or out of place at all: Of course living things are constantly doing things, engaging in various actions and processes that keep them alive, such as respiration and digestion and so on. That's just how living things work, and we can see those workings in ourselves as well as in the organisms that surround us.

[25] Actually, the accusation that Aristotle thought of non-living things in these terms seems entirely based on a series of misunderstandings and mistranslations, or so Joe Sachs persuasively argues. [Sachs, see especially p.13-17]

So Aristotle maintains that the *psyche* is the active, self-sustaining wellspring of a living thing's continued bodily existence. But what is it, really? Let us return to *On the Soul* for a fuller answer:

> The soul is the cause or source of the living body. The terms cause and source have many senses. But the soul is the cause of the body in all three senses which we explicitly recognize. It is the source of movement, it is the end, it is the essence of the whole body. [Aristotle, *On the Soul*, p.661 (415b, 9-11)]

One need not be an Aristotle scholar to recognize his familiar doctrine of causes in this quotation. "The source of movement" is 'efficient cause,' the only one of Aristotle's four causes consistent with the typical modern understanding of the word "cause." "The end" is 'final cause' or *telos*, the completion or goal of an activity. "The essence" is the 'formal cause,' which is the complete form of the thing – the shape, the organization of its parts, etc.[26] If the 'formal cause' (*eidos*) is the shape and the organization of parts and so on, then *psyche* is not such a mysterious and troublesome concept after all. A living thing has various parts, organs: *Psyche* is, among other things, the specific arrangement and activity of the organs that allows the animal to keep on being what it is.

Aristotle goes on to say that the formal cause/*eidos* and the final cause/end/*telos* are both central to the meaning of *psyche*:

> That it [soul] is the last, is clear; for in everything the essence is identical with the cause of its being, and here, in the case of living things, their being is to live, and of their being and their living the soul in them is the cause or source...
>
> It is manifest that the soul is also the final cause. For nature, like thought, always does whatever it does for the sake of something, which something is its end. [Aristotle, *On the Soul*, p.661 (415b, 12-16)]

[26] 'Essence' is another problematic translation. Aristotle uses the word *eidos* in this passage, which James Lennox argues is properly translated as 'form' and never as 'essence.' The concept of 'essence' was not captured precisely by any single word in Greek, although Aristotle indicated the concept more or less consistently using the phrase 'what it is to be' (*to ti ên einai*). [Lennox, p.129]

In other words, the form (*eidos*) and end (*telos*) are inseparable in the case of living things because the end of a living thing is its own continued existence, and its form determines its existence as the kind (*genos*) of thing that it is. As discussed previously, the latter is the element of Aristotle's thought that was such an obstacle for the progress of biology, for it is the conceptual heart of the idea of species as fixed: On this view, each kind has a fixed, immutable form/*eidos* passed on from parent to child[27], and all variations are inessential and incidental.

This makes a certain amount of sense in context, for Aristotle was primarily interested in nature's generalities, categories and unities, not nature's specificity and diversity. In contrast, modern evolutionary biologists – indeed, all scientists – seek to account for nature's generalities by understanding the underlying specifics: Whereas Aristotle thought of individuals merely as examples of an unvarying, underlying species/type, one of the primary insights of Darwin was to see species as a population of unique individuals. Thus, the first modification of Aristotle's concept of the soul we must make to bring an Aristotelian conception of life more in consonance with modern evolutionary biology is to reinterpret the phrase 'being what it is' in an individual rather than generic fashion, to change it from 'being the kind of thing that it is' to 'being the specific thing that it is' – more simply, to 'being itself.' But we can shed this objectionable essentialism in the concept of a species or kind without abandoning the teleological core of Aristotelian biology that is important for my purposes here.

Changing Aristotle's conception of *eidos* from kinds to specifics requires a similar alteration in the conception of *telos*. For Aristotle, the *telos* of a living thing is 'to keep on being

[27] This is not wholly accurate. Rather predictably given the status of women in Greek culture, it was taken for granted by Aristotle that the form/*eidos* was passed on from the male parent, and that females provided only the matter (*hulê*) on which the form was imposed. *Hulê* was Aristotle's fourth cause, usually translated as 'material cause,' and simply means that which underlies form, the stuff of which a thing is made. Aristotle's conception of matter was very different from the modern conception, for his matter was a sort of neutral substrate with no particular properties of its own. It is not particularly relevant to the matters at hand in any case, so I will say no more than that.

the kind of thing that it is' – that is, to actively maintain its *eidos*. Now that we have seen how the facts of biology require us to construe the *eidos* of a living thing individually, its *telos* must be similarly individual. A living thing's *telos* is to keep on being itself, not to keep on being a thing of its kind. A living thing's *telos* is its own unique survival, not the survival of its kind or its survival simply as a representative of its kind.

However, my prior analysis of the conceptual core of evolution by natural selection suggests we take one further step away from Aristotle's understanding of life. From the modern perspective, survival is still not quite the right *telos*: A living thing merely being itself has no more bearing on natural selection than its merely being a member of its kind; it must also reproduce itself, for natural selection consists in differential reproductive success over generations. It is not merely sustaining its individualized *eidos* that is the *telos* of every organism, but passing on its *eidos* to succeeding generations.

It may be tempting here to read too much poetry and metaphor into this reinterpretation of Aristotle's biology within an evolutionary framework, by which I mean something along these lines: Every organism strives to pass on its *eidos*, its unique essence, its genome. This would be a mistake. It is no more accurate to view Aristotle's *eidos* as a genome (or possibly something more abstract, like the information embodied in a genome) than it would be to view Aristotle's *eidos* as an eternal, unchanging, and wholly abstract Platonic form (also the word *eidos*, but used to indicate a rather different idea). Aristotle's *eidos* is much more concrete, consisting in the complete arrangement of parts manifest in the actual bodies of a given kind of living things. My evolutionary reinterpretation of Aristotle makes *eidos* even more concrete, leaving out 'kinds' entirely and conceiving of *eidos* as the complete arrangement of parts of the body of each particular living thing. Thus, this reinterpreted *eidos* most closely parallels the phenotype rather than the genotype of an organism, if any such parallel is to be drawn – and it would be better still to leave off all talk of 'types' entirely and stick to particulars so as to avoid confusion.

While the claim that the *telos* of an organism is its own reproduction does mean that an organism's *telos* is the reproduction of its *eidos*, it also means that the *telos* of an organism is the reproduction of its *telos*: Reproduction of the whole organism is at stake, not just its *eidos*, because the *eidos* is not separable from the *telos* and the body. If I string together the segments of Aristotle's *On the Soul* which I cited above, which appeared more-or-less consecutively in that text but which I separated for the purposes of analysis, the following equivalences follow in rapid succession: The soul (*psyche*) and the body are one, and the form (*eidos*) is the soul, but the end (*telos*) is also the soul, so the form and the end are inextricably one with the soul, and thus with the body. Meaning what, exactly? Remove the obstructions of ancient abstractions, and include the reinterpretation of *eidos* and *telos* necessary to be consistent with what we know about organisms that Aristotle did not, and all it amounts to is this: The physical arrangement of a living thing's body – more colloquially and simply, the way it's put together – and the activities that the body engages in are as they are *for the sake of* its reproduction.

The teleological notion of being 'for the sake of' some end neither implies nor requires any intention or awareness, even on Aristotle's original view: Plants, for example, have neither intentions nor awareness, but still have a *telos*. Nor is there anything opposed to a mechanistic understanding of nature in such a teleological understanding of biology, not even Aristotle's version. Aristotle's 'soul' consists in the arrangement and operation of the various parts of an organism's body – for the sake of sustaining itself in Aristotle's version, for the sake of reproducing itself in my revision. There is nothing about such bodily arrangements and operations that requires or implies anything but ordinary step-by-step causal processes in either version.

Perhaps the most significant change introduced by my reinterpretation of Aristotle's teleological understanding of biology is the explanation of how organisms come to be structured around a *telos*: As Mayr noted (see p.78), Aristotle could conceive of no reason why organisms' parts should be so arranged except chance or purpose – and it clearly is not chance, hence Aristotle invoked what Mayr called a 'cosmic teleology' to explain how organisms came to be the way they are. One way to understand Darwin's great insight is that he saw false dilemma of Aristotle's "chance or purpose" dichotomy; he realized that chance and necessity combine to produce all the results that were previously be attributed to purpose. In fact, Darwin saw that chance and necessity were much a more plausible cause, being much more consistent with the actual messy, wasteful, imperfect, far-from-harmonious natural world than with any purpose ever proposed.

Likewise, my evolutionary reinterpretation of Aristotle provides a ready account of how this teleological structuring comes to be – the same account as Darwin, in fact, but rephrased in different terms: Bodies whose form and activities are structured for the sake of reproduction are inevitably produced by the iterative operation of natural selection's ruthless filter over generation after generation. Organisms whose forms and activities are less well-suited to reproduce themselves (within whatever constraints a given population of organisms faces) are represented in diminishing proportion in successive generations, and those aspects of their forms and activities which are very poorly-suited may eventually be eliminated from a population entirely. Organisms whose forms and activities are better suited to reproduce themselves are represented in greater proportion in successive generations, and those aspects of their forms and activities which are very well-suited may eventually be present in every member of a population.

Thus, given the modifications I specified – abandoning the 'member of its kind' conception of an organism for a conception that recognizes the individual nature of each organism's *eidos* and *telos*, and recognizing that reproduction rather than mere continuance/survival is the true *telos* of each organism – Aristotle's understanding of the essence of life would seem to be entirely compatible with modern biology. However, it isn't particularly useful to the project at hand to keep using the terms like 'soul' (*psyche*) and 'form' (*eidos*) with all their attendant historical and philosophical baggage, so I will abandon them.

Telos, however, is another matter. This is exactly what I have been working towards; I sought a teleological conception of life not merely consistent with evolutionary biology, but thoroughly rooted in it. The motivation for doing so was straightforward: Aristotle's conception of *telos* does not merely imply or define value, it determines what is valuable to that which has a *telos*. A being's *telos* is its highest good, that which is of intrinsic value to it, that by virtue of which every other valuable state of affairs has (extrinsic) value for that being.

A *telos* is also a matter of fact: To say "*A* has *T* as its *telos*" makes a claim about *A* that is true or false, not itself contingent on any prior value claim. Determining that a being has a *telos* is thus the perfect fact-value bridging normative claim, for a being's *telos* is the standard by which the truth of other value claims can be determined: "Is state of affairs *X* valuable to *A*?" becomes a question that can be answered by determining whether and how *X* contributes to *A* achieving its *telos*, or highest good. This is why Aristotle's claim about the human *telos* served as the key fact-value bridging foundational claim for his virtue theory: If Aristotle's claim that happiness is the human *telos* is false, his entire approach to value fails and his ethical theory is based on nothing at all.

I believe I have established that reproduction is a *telos* in an Aristotelian sense (albeit not a *telos* he identified as such), and hence an organism having reproduction as its *telos* constitutes a claim about what is of **intrinsic** value to that organism. And given that an 'objective value claim' as I define it (see p.26) is simply a factual claim about what is of value – or if not a straight factual claim, at least a value claim that is fact-like insofar as it is capable of being assessed as true or false – my claim that the *telos* of every organism is its own reproduction is an **objective** value claim. However, the *telos* of reproduction is **not universal** – except in the loose sense that every organism has such a *telos* – because what is valuable to each organism is not the same state of affairs: It is not reproduction in some abstract sense that is an organism's *telos*, but rather *its own reproduction*. Thus the *telos* of reproduction is clearly **proprietary**, and so does not fulfill the criteria necessary for it to serve as the foundational fact-value bridging claim I seek.

On the other hand, the reproductive *telos* still serves as a fact-value bridging normative claim of sorts: It constitutes a normative value claim which can be used as a basis for reaching prescriptive conclusions, but it provides only a prudential norm, not a moral norm. If an organism's *telos* is its own reproduction, that *telos* generates a nested hierarchy of those states of affairs (maintaining homeostasis, finding food, avoiding predation, securing a mate, etc.) which are of value to that particular organism insofar as they contribute to its reproduction. In other words, the reproductive *telos* provides a naturalized account of pragmatic/instrumental value, if not an account of moral value: The reproductive *telos* determines what an organism *ought to do* in its own best interest, and defines a meaningful sense in which any organism from a bacterium to a baboon can be said to have a best interest.

To return once more to the theme of solving Hume's fact-value problem, these ought claims generated by the reproductive *telos* are a new kind of relation or affirmation we can make about organisms, and this new relation has indeed been "observ'd and explain'd," and reasons have been given for it. What reasons? Insofar as my claim 'the *telos* of every organism is its own reproduction' is simply a way of restating the consequences of natural selection, these (non-moral) value claims are justified by the same inductive arguments which support the theory of evolution by natural selection, and are based upon the same empirical evidence. Thus, the *telos* of reproduction is as thoroughly justified a fact-value bridging normative claim as I could desire. But insofar as the norm at stake is proprietary to each individual organism, the *telos* of reproduction provides but an exceedingly narrow bridge between fact and value: The next step in my argument will be to determine whether and how this bridge might be broadened.

CHAPTER 4

NON-PROPRIETARY EVOLVED VALUE

Section 1: Understanding the reproductive *telos*

In my experience, even people who understand and accept evolution in general and human evolution in specific sometimes balk at the idea that we too belong in the category of organisms whose *telos* is their own reproduction. I find this rather strange, since these same people generally have no problem with the idea that distinctive features of our species – a large prefrontal cortex, bipedalism, tool use, language, culture and so on – became features of our species in the first place because incremental improvements in such traits helped our ancestors survive and reproduce themselves. So what is the basis for this balkiness?

At first I was tempted to think this just another example of lingering human exceptionalism, the desire to see humans as separate from the rest of nature: Even those who don't truly believe we were created often have trouble not seeing humanity as a special creation, as fundamentally different from other animals. While I won't deny that such attitudes have a great deal of sociocultural inertia, I have come to believe that the most common basis for rejecting the reproductive *telos* is genuine misunderstanding, not just a visceral reaction (although in most instances there is some of the latter as well). Such misunderstandings are worth examining directly to clarify what the reproductive *telos* actually implies, and to distinguish that from what it might seem to imply.

The most common form of objection I've encountered treats the reproductive *telos* as a straightforward universal claim and attempts to refute it by citing exceptions, always human exceptions. Such a response usually takes a form along these lines: "But many people don't even want children. Heck, *I* don't want children! So it can't be true that reproduction is the *telos* of *every* organism." This kind of objection might be based on any number of the following misunderstandings.

The most obvious misunderstanding is based on confusion about what sort of claim the reproductive *telos* is. The most important claim is the normative claim, the claim that an organism's own reproduction is intrinsically valuable to it: A claim about what is of intrinsic value to an entity is the basis both for determining what is extrinsically valuable to that entity and for drawing prescriptive conclusions for that entity. However, even when talking about humans, saying that some state of affairs is valuable to someone does not necessarily imply that this someone perceives it to be valuable or knows it is valuable. After all, the entire point of my attempt to establish an objective normative claim is to make it possible to say that someone is mistaken in what he or she takes to be of value, and why (see p.26). Thus, to say that state of affairs *T* is your *telos* is not to say that you want to bring about *T*, or even that you ought to want to bring about *T*: Rather, it is to say that *T* is of intrinsic value to you whether you know it or not, whether you subjectively value it or not.

Now one might object here that the concept of *telos* implies more than that, and indeed it does: To say that any given *T* is the *telos* of an entity implies that some significant proportion of the entity's characteristics and activities are for the sake of *T* – that is, organized around and directed towards *T* as an end. Speaking colloquially, shouldn't this imply that these big evolved brains of ours should be focused on and directed towards our own reproduction? And wouldn't the fact that some humans don't even want children then be a counter-example to that claim?

My response to such an objection highlights a second misunderstanding of the reproductive *telos*: To say that the *telos* of every organism is its own reproduction does not imply that every organism does successfully reproduce itself. Rather the opposite, given that the origin of the reproductive *telos* is natural selection, and natural selection operates precisely because every organism does not successfully reproduce itself – and only those that do succeed are represented in future generations. Thus, saying that the *telos* of reproduction is universal does not mean that reproduction itself is universally realized: To say that the *telos* of the acorn is to become the oak so as to produce more acorns (potential offspring) does not imply or require that every acorn does indeed successfully become an oak. It is simply a fact that the reproductive strategy of the oak tree involves the massive overproduction of acorns, which are after all just embryonic oak trees.

Similarly, even if the physiological basis for the human capacity for massively facultative behavior[28] – those "big evolved brains of ours" – sometimes lead a human to reproductive failure rather than success, that capacity is still in us because of its contributions to our ancestors' reproduction, and that capacity will remain in us as long as it leads to continued future reproduction significantly more often than it leads to non-reproduction. It is simply a fact about our reproductive strategy that it depends on a great deal of behavioral flexibility, and that range of possible behaviors includes many behaviors that result in humans not reproducing themselves (although not in the same percentages as acorns that do not become oaks).

To put the same point less precisely but more clearly, big brains that on occasion decide not to have kids are better (for reproduction) than smaller brains that aren't as good at all the other things brains do which contribute to humans survival and reproduction on the vast majority

[28] Facultative behaviors are those which an organism changes in response to different environmental conditions.

of occasions. Generally speaking, success matters more than failure in evolution: If a given human organism does not leave offspring, it doesn't matter from an evolutionary perspective whether that failure to reproduce results from a conscious lifestyle decision or a tragic hedge trimmer accident – what matters is the traits passed on by those who do successfully reproduce.

Another confusion lurks within this particular line of argument, because it seems to imply that the intrinsic value of our own reproduction is the only normative claim at stake in the reproductive *telos*. However, it is important to remember that any number of things might be extrinsically valuable based on their contributions to achieving an intrinsically valuable state of affairs. Consider the characteristically human activities even people who choose not to have children engage in: Our careers allow us to secure resources, our social engagements build communities of mutual support and obligation, and most of us pursue romantic relationships and domestic partnerships even if we lack any intention of child-rearing. For humans, resource-gathering, community-building and intimate relationships are behaviors extrinsically valuable for the sake of their contributions to the self-reproduction that is intrinsically valuable to every organism. Thus, even those who choose not to reproduce typically put their time and energy into other activities which fit very well into the nested hierarchy of ends generated by the reproductive *telos*, which does nothing to belie the expectation that some significant proportion of the entity's characteristics and activities are focused on and directed towards that *telos*.

But perhaps the most transparent confusion of an objection along "But I don't even want children!" lines is taking the reproductive *telos* as implying some sort of negative judgment about that choice. In response, I can only repeat that the *telos* of reproduction is strictly proprietary and thus not a proper basis for any sort of ethical theory, so any prescriptive conclusions which might follow from it don't have the overriding character of a moral claim. It may be that a person who chooses not to have children is failing to do what is in his or her own biologically determined self-interest, but there is no implication that he or she is *wrong* in failing to do so. In fact, that would be decidedly odd: When we refer to the overriding character of moral prescriptive claims, what is it that they override? Typically, moral prescriptions are taken to trump the prescriptions of self-interest – and not just any prescriptions of self interest, but specifically those which ignore the interests of others entirely, the feature that distinguishes genuine selfishness from merely acting in one's own self-interest.

In this typical opposition of morality and selfishness lies a clue for moving forward. From the perspective of biology, what does it mean to take the interests of other organisms into account? The interests at stake in this discussion are reproductive interests, so we must ask whether and how some organisms might have an interest in the reproduction of organisms other than themselves. This is not a new question, for various sorts of cooperative and even altruistic behaviors are apparent in all sorts of organisms from bacteria living in colonies to social mammals alerting each other to danger – and in our own behavior, of course. An understanding of the evolution of social behavior may reveal ways in which the reproductive *telos* is not as strictly proprietary as it at first appears.

Section 2: Altruism, kin selection, and a less proprietary reproductive *telos*

> Would I lay down my life to save my brother? No, but I would to save two brothers or eight cousins.
> – attributed to J.B.S. Haldane by John Maynard Smith from "a pub discussion" [Lewin, p.325]

Suppose, in what I can unfortunately assure you is a wholly imaginary example, that a rich and childless (for whatever reason) maternal uncle has recently given me, no strings attached, a great deal of money – more than enough to secure the material needs of myself and my future children for decades, to secure the best available health care and education for all my children and even my eventual grandchildren, and so on. Suppose further that I would never have gotten around to having children at all without my uncle's generosity – perhaps thinking that I should pay off my student loan debt first, and subsequently deciding that age 70 was no time to start a family. Surely, then, my hypothetical uncle can be seen as having advanced my reproductive interests substantially with his generous gift.

On the other hand, perhaps my uncle has not so much done something for me as he has done something for himself, insofar as advancing my reproductive interests also advances his own reproductive interests – at least from a somewhat oversimplified and gene-centered evolutionary perspective: Roughly one-half of my mother's DNA is identical to my maternal uncle's DNA, and roughly one-quarter of my own DNA is identical to his. By making it much more likely that I have children and grandchildren, my uncle is advancing his own reproductive interests by half as much as he would have by similarly aiding his own child – and since he is by stipulation childless and likely to remain so, half a child's worth of reproductive success is better than none.

So has my uncle simply advanced his own good as defined by his proprietary reproductive *telos*? Or is the fulfillment of my reproductive *telos* of genuine value to him, thus making his own reproductive *telos* just a little less proprietary (if not genuinely non-proprietary)? Do I have to choose between these alternatives, or can both be true? Answering these questions will require, at the very least, digging a little further into those aspects of evolutionary biology the example is intended to illustrate.

First, a bit of vocabulary: **Fitness** is simply a measurement of reproductive success, often calculated for an individual organism by simply adding up the total number of the organism's offspring and/or grand-offspring which survive to reproductive maturity. Heritable traits which increase fitness become more widely represented in successive generations, and traits which decrease fitness become less represented – which is just a more streamlined way to describe the process of natural selection, already described in several different ways in the preceding chapter.

Kin selection is the theory that natural selection will not only favor heritable traits which increase an organism's own fitness, but also traits which increase the fitness of an organism's relatives – possibly even if they do so at the expense of the organism's own fitness: Mathematically, a trait which has a negative impact on fitness can still undergo positive selection (become more widely represented in successive generations) if the negative fitness impact of the trait on an individual organism is balanced by a positive fitness impact on other organisms which share the same trait. If a given trait is heritable (which it must be to undergo selection in the first place), an organism's relatives have a greater probability of also having and passing on that trait than a randomly selected member of the population – and the closer the relation, the higher the probability. Hence kin selection, whereby a trait which has a negative impact on an individual organism's fitness spreads through a population in spite of that because of its mathematically greater positive impact on the fitness of an organism's relatives. The logic of kin selection is loosely described in my imaginary uncle example, and is used to humorous effect in the J.B.S. Haldane quip I quoted to open this section.[29]

Kin selection was proposed not out of purely theoretical considerations, but to solve a particular problem, which might be dubbed the **problem of altruism**: Insofar as natural selection is an engine driven by individual reproductive success (which is measured by calculating fitness), cooperative behavior in which individual organisms decrease their own fitness to

[29] A detailed account of the theoretical and mathematical foundations of kin selection/inclusive fitness can be found in W.D. Hamilton's "The Genetical Evolution of Social Behaviour" [Hamilton, 1964].

increase the fitness of other organisms seems inexplicable. In other words, natural selection for fitness-sacrificing behavior seems self-contradictory on the face of it. Kin selection is just an explanation of how the process of natural selection can account for such behaviors when one looks past the face of it – and kin selection has been shown empirically to account for many otherwise inexplicable behaviors evident in social arrangements from bacterial colonies to beehives. Because various degrees of social behavior are such widespread phenomena in nature, some biologists have argued for relying more on the measure of **inclusive fitness** which incorporates both an organism's offspring (individual fitness) and the offspring of relatives in proportion to their relatedness (using various calculating methods). Measuring inclusive fitness instead of or in addition to individual fitness articulates kin selection more clearly as just one way in which natural selection operates, rather than as a special type of selection in its own right.

Since an organism's inclusive fitness is still properly its own fitness in some broadened sense, some biologists have argued that kin selection explains only the appearance of altruism and cannot account for behaviors which are "genuinely" altruistic, thereby re-defining 'altruism' to exclude sacrifices of fitness which primarily benefit relatives [Trivers, 1971, p.35]. On such a view, only behaviors which decrease an organism's fitness (however calculated, even inclusive fitness) and increase the fitness of unrelated organisms[30] can be termed altruistic. Various mechanisms operating within the framework of natural selection have been proposed to explain

[30] Of course, no earthly organisms are actually "unrelated" *per se*. "Unrelated organisms" means, in this context, all those members of a population whose alleles are identical by descent to those of a given member of the population to a degree not significantly greater than chance. "Alleles" are the paired forms of a given gene present at a given locus of an organism's DNA; sexually reproducing organisms receive one allele from each parent (except on the sex-determining chromosome). The term "identical by descent" is used in contradistinction to alleles which are "identical by type," those alleles which produce the same phenotypic effect even though they do not have the same sequence of codons: For example, there could be multiple forms of a dominant allele for black fur in cats, so two cats might both have black fur even though they have different inherited codon sequences in their dominant fur-coloration alleles – so those alleles would be identical by type rather than identical by descent.

how such "genuinely" altruistic behaviors can indeed evolve and have evolved, such as **reciprocal altruism** and **group selection** (the latter of which is especially controversial).

However, it is not clear that even behaviors explained by reciprocal altruism and group selection are "genuinely" altruistic: Since even these mechanisms operate with the framework of natural selection, they must in some way or other advance the proprietary reproductive *telos* of the organisms who engage in the (allegedly) altruistic behavior. I will call this doubt that any solution to the problem of altruism actually makes the reproductive *telos* any less proprietary the thesis of **evolutionary selfishness**.

By this point, the thesis of evolutionary selfishness, the problem of altruism, and the controversy surrounding its potential solutions may sound vaguely familiar even to those with no knowledge whatsoever of evolutionary biology: The judgment that my imaginary uncle only *appears* to be doing something altruistic but is *really* acting in his own (evolutionary) self interest has many parallels with the pervasive belief generally referred to as **psychological egoism**, perennial fodder for introductory philosophy classes. Because psychological egoism involves positions and arguments which are familiar to a broad audience, I will use it as an illustrative analogy for addressing the problem of altruism and the evolutionary selfishness thesis.

Psychological egoism purports that every person's every action is best explained by that person doing what they perceive to be in their own best interest (or some close variation on that definition). From the perspective of psychological egoism, even the most apparently selfless act of generosity is motivated by the actor's self-interest – for doesn't such an action make the actor feel good, and/or look good in the eyes of others? Taken to its extreme – as it commonly is – psychological egoism even maintains that the most secret act of generosity in which the giver takes no overt joy is still explicable in terms of self interest: Perhaps such an act confirms the giver's self-image as a good person, thus serving his or her self interest even if he or she takes no particular pleasure in the act. Although not necessarily committed to psychological egoism himself, Kant took the inability to rule out self-interested motives very seriously:

> [I]f we pay attention to our experience of what human beings do and fail to do, we encounter frequent and, I must admit, justified complaints that one cannot in fact point to any sure examples of the disposition to act out of pure duty. Thus we hear the charge that, although many things may be done that are in accord with what duty commands, it still remains doubtful whether those actions are really done out of duty, and doubtful therefore whether they have moral worth. That is why there have always been philosophers who absolutely denied the reality of this disposition in human conduct and ascribed everything we do to more or less refined self-love... It is in fact absolutely impossible to identify by experience, with complete certainty, a single case in which the maxim of an action – an action that accords with duty – was based exclusively on moral reasons and the thought of one's duty. There are cases when the most searching self-examination comes up with nothing but duty as the moral reason that could have been strong enough to move us to this or that good action or to some great sacrifice. But we cannot conclude from this with certainty that the real determining cause of our will was not some secret impulse of self-love, disguising itself as that Idea of duty. So we like to flatter ourselves with the false claim to a nobler motive but in fact we can never, even with the most rigorous self-examination, completely uncover our hidden motivations. [Kant, p.208, (4:406)]

If we re-word Kant's concerns to something more in line with the usual consequentialist discussions of psychological egoism, his analysis suggests it is absolutely impossible to claim on the basis of experience that any action is genuinely altruistic – motivated by the desire to do good (for someone else) rather than by self-interest. Even introspection cannot give us reliable evidence of altruistic motive, because we do not have perfect access to our own motivations: Some of our motivations are hidden, even from ourselves. Interpretations of our actions consistent with psychological egoism are always possible: Even if they are unlikely or implausible, they cannot be eliminated as potentially true, as the real motivations of our secret selves.

This inability to eliminate an egoistic explanation for any action whatsoever makes psychological egoism quite rhetorically persuasive, but practically useless. If a theory can offer an explanation for every possible phenomenon within its scope, even directly opposing phenomena such as a given individual performing or refraining from the exact same action on a given occasion, then two consequences follow: First, the theory which originates such explanations cannot be falsified by any possible evidence – so neither can it be tested or verified in any way. What would constitute a test for psychological egoism? Expecting to find self-interest motivating an action and being able to find none. But when psychological egoism proposes hidden motivations such that even the actor cannot be certain that her or his motivation is not subconsciously self-interested, it rules out every possible test in advance.

Second, such a theory has no explanatory value whatsoever: At the very least, an explanation should tell us why this particular event occurred instead of some other possible event. Psychological egoism never explains why action *A* happened instead of *not-A*, because it gives exactly the same explanation for both actions: If an actor does *A*, it's because *A* was the action the actor perceived, consciously or not, to be in his or her own best interest in that circumstance. And if an actor refrains from doing *A*, it's because *not-A* was the action the actor perceived, consciously or not, to be in his or her own best interest in that circumstance. With no way to access an agent's entirely presumptive hidden motivations, there is no genuine explanatory content either way.

Even if the unfalsifiable nature of psychological egoism means that we have no genuine reason to believe it is true or find it useful, we may still have reason to be worried if some other conclusion we wish to advance requires psychological egoism to be false. The universal claim of psychological egoism directly opposes an existential claim I might as well dub 'psychological altruism' – that is, the claim that one or more actions by some individual at some time has been or might be motivated solely by a genuine desire to advance the interests of another rather than being motivated by self-interest, overt or covert. If no conclusive evidence can be advanced to falsify psychological egoism, then no conclusive evidence can be advanced to support even a single instance of psychological altruism – hence Kant's judgment that it is impossible to make an empirical case that anyone has ever actually acted from a purely moral motivation.

The objection that kin selection doesn't account for "genuine" altruism seems very much like the objection against psychological altruism rooted in psychological egoism: Both objections seem to presume the impossibility of altruism in advance, even to the point of reinterpreting the evidence to preserve the anti-altruism position – in one case by positing hidden psychological motivations, and in the other case and by redefining 'altruism' itself. It seems clear that positing hidden psychological motivations is an *ad hoc* premise with no justification beyond saving the psychological egoism thesis. Do various attempts to re-define evolutionary altruism have the same *ad hoc* character?

On one hand, it would not seem so: Re-defining 'altruism' to make a distinction for the purposes of research is perfectly legitimate. That is, if one has a research interest in how altruistic behavior towards non-kin organisms can evolve, then it makes sense to define 'altruism' as behavior that sacrifices an individual's own fitness to advance the fitness of non-kin members of the individual's population. On the other hand, cooperative behavior between organisms is no less cooperative behavior whether it occurs between close kin (as in a haplodiploid bee colony) or between sometimes "completely" unrelated individuals (as in a chimpanzee troupe). If the phenomenon one cares about is the evolution of cooperation – which, after all, is where the problem of altruism arises – then narrowing the definition of 'altruism' to exclude what biologists broadly agree is the most common form of selection for cooperative behavior (kin selection) seems somewhat question-begging at best.

The problem of altruism, whether evolutionary or psychological, is intimately related to the question of proprietary versus non-proprietary value. Psychological egoism seems to imply that humans always value other humans and their welfare in a strictly proprietary manner. Similarly, evolutionary selfishness seems to imply that the value of the reproductive *telos* is always strictly proprietary. Thus, the Kantian concern discussed above seems just as relevant to the conflict between the evolutionary selfishness thesis and the various solutions to the problem of altruism as it is to the conflict between psychological egoism and psychological altruism: It is not enough to point out that there is something question-begging and *ad hoc* about objections to the evolution of altruistic behavior, because making the case that the reproductive *telos* is not strictly proprietary would seem to require ruling out the evolutionary selfishness thesis. Thus, my next task is to determine whether and how evolutionary selfishness might be ruled out even though psychological egoism apparently cannot be. As a first step, it is worth asking whether or not psychological egoism is really as difficult to rule out as it appears.

Actions motivated by concern for another's welfare are generally taken to serve as counterexamples to the universal claim of psychological egoism. The defender of egoism dismisses these examples by saying that they are primarily motivated by the desire to feel good about one's altruistic action, or to gain the social benefits of being seen as an altruist, rather than being motivated by concern for another's welfare. In cases where the altruist cannot discern even in himself or herself the slightest pleasure in the altruistic act, and where there is no conceivable social benefit to oneself (as with a secret act of altruism), defenders of psychological egoism posit hidden motives of which even the altruist is unaware. This claim of hidden motives seems entirely *ad hoc*, since no one can produce evidence for these motives if they are hidden even to the one moved by them. As such, this argument only preserves the bare possibility that psychological egoism *might* be true in spite of counterexamples, on the basis that these might only *appear* to be counterexamples.

I would argue, however, that the defender of psychological egoism's argument against counterexamples fails before they resort to hidden motives. Even if an altruistic action does make the altruist feel good or look good in the eyes of others, that argument does not undermine the claim that the altruist is motivated by the altruistic aim of advancing another person's welfare. At best, citation of these other motivations (pleasure, social approval, etc.) makes a case that the altruist might always *also* have one or more self-interested motivations, in addition to the selfless motivation of advancing another's welfare: Simply noting the existence of additional self-interested motives does nothing to make the case that the altruist's *primary* or *only* motivation is those other, self-interested motives. As such, the thesis of psychological egoism – that every person's every action is *best explained* by that person doing what they perceive to be in their own self-interest – is not really defended from these sorts of objections by counterexample. If the best defense of psychological egoism on offer is that there might also be a self-interested motive even when the only apparent motives are selfless, the truly best explanation of altruistic actions as a category would appear to be that altruists are moved to act by altruistic motives and possibly also by self-interested motives. The truth of psychological egoism's universal claim about the motivations of human action thus stands refuted by the existence of selfless motives even if it is impossible to rule out the presence of self-interested motives in addition to selfless ones.

Similarly, the argument that an altruistic behavior advances an organism's own reproductive *telos* does not in any way undermine the fact that the behavior also advances the reproductive *telos* of another organism. Analogy would seem to suggest that a case can be made for a non-proprietary component of the value of another organism's reproductive interests even if an additional proprietary component cannot be ruled out. In other words, there is nothing genuinely contradictory in concluding that the value determined by the reproductive *telos* can be proprietary in one respect and non-proprietary in another respect at the same time.

However, it is worth considering whether the analogy may break down because of this important difference between analogues: The thesis of psychological egoism focuses entirely on motivation, whereas the evolutionary selfishness thesis, framed in terms of the reproductive *telos*, focuses on value. While value is surely related to motivation, they cannot be treated as equivalent. To understand how the difference might affect the analogy, it is necessary to discuss the relation between motivation and value.

Recognition or perception of something as valuable – without regard to whether what is perceived as valuable is of genuine objective value – is generally taken to be motivating, although it need not provide decisive motivation that results in action. Thus, although psychological egoism is usually framed in terms of motivation, it clearly has implications for value. Specifically, the debate between psychological egoism and altruism implies a conflict over whether we value the welfare of others **intrinsically** or **extrinsically**: Is an altruistic action motivated by the perception that another's welfare is intrinsically valuable, or is an altruistic action motivated by the perception that the welfare of another is extrinsically valuable to advancing the alleged altruist's own (intrinsically valuable) welfare? My argument that psychological egoism can be refuted is unchanged, but that argument no longer hinges on the issue of whether the value of others' welfare is non-proprietary, but rather on whether it is intrinsic.

The analogy between psychological egoism and evolutionary selfishness is thus exposed as being fraught with potential confusions between the proprietary/non-proprietary value distinction and the intrinsic/extrinsic value distinction: If the egoism vs. altruism debate is really about whether the welfare of others is intrinsically or extrinsically valuable, perhaps all this analogy suggests is that mechanisms like kin selection do nothing to make values determined by the reproductive *telos* less proprietary: Rather, kin selection and similar mechanisms may simply show how the reproductive interests of other organisms can be extrinsically valuable to an altruist by advancing its intrinsically valuable reproductive interests. I think this suggestion is misleading, however; seeing how and why it is misleading requires examining exactly how the intrinsic/extrinsic and proprietary/non-proprietary value distinctions relate to one another.

Whether or not something has value in a non-proprietary fashion hinges on whether its value is limited to a single valuer: To return to the discussion of Mill's argument for the principle of utility where I first made the proprietary/non-proprietary distinction (see p.40), it is not quite correct to say that every person only values his or her own happiness. Rather, people value happiness – they see happiness as desirable; and because they see it as desirable, they are motivated to pursue it. The motivation to pursue happiness may be dependent on whose happiness it is – my happiness motivates me, your happiness motivates you, and therefore is proprietary – but happiness itself is of value to any valuer in a non-proprietary fashion. However, motivation springs from perception of value, so it would be more accurate to say that the value of happiness has both a proprietary and non-proprietary component – the non-proprietary component being the foundation for the principle of utility.

In contrast, the intrinsic/extrinsic value distinction is not about how value is related to valuers, but whether or not the value of a state of affairs depends on the value of some other state of affairs. Sticking with utilitarianism to complete the parallel, happiness is intrinsically valuable because it is valued for its own sake, not for the sake of achieving something else of value.

However, consider the following: Extrinsic and intrinsic value are not mutually exclusive (see footnote on p.23), for one and the same state of affairs can be valuable in itself and also valuable for how it contributes to something else of value. For example, if you are a friend or loved one, your happiness may contribute to my happiness. Your happiness in such a case would still be of intrinsic, non-proprietary value, but your happiness would also be of extrinsic value *to me*. The "to me" component cannot be ignored: The extrinsic value of your happiness, insofar as it is extrinsic, must also be proprietary. Why? Because the intrinsic component of the value of your happiness is also the non-proprietary component that grounds the principle of utility – happiness as an end in itself, no matter whose happiness it is. The extrinsic component is found specifically in the contribution of your happiness not just to happiness in the abstract, but in its contribution to *my* particular happiness. Logically, this must be the case whenever something is of both intrinsic and non-proprietary value: Any extrinsic value it has in addition to that intrinsic, non-proprietary value must always be proprietary.

This logic is also reversible: If some state of affairs (my friend's happiness) is of extrinsic, proprietary value, that means it has value for its contribution to something else of value to a particular valuer (my happiness). If the something else (my happiness) weren't of value *to a particular valuer* (if it weren't *my happiness* in particular, but happiness generally), there would be no reason to call this extrinsic value "proprietary." If that same something (my friend's happiness) also has intrinsic value, the intrinsic component of its value does not depend on that something else of value for a particular valuer (my happiness), so the intrinsic component of its value would also be non-proprietary.

Now that the way non-proprietary/proprietary and intrinsic/extrinsic value interact is more clear, I will return to the analogy between psychological egoism and evolutionary selfishness, starting with egoism. To see that the psychological egoism thesis is indeed refuted, it is worth restating the thesis and summarizing the argument against it.

However it is formulated in detail, every version of the psychological egoism thesis states that the best (or "real" or "ultimate") explanation of and reason for all human actions is that we are motivated by self-interest. On the common understanding of the connection between motivation and value, being motivated by self-interest means that we are motivated by perceiving something to be of value to us – that is, as having proprietary value. Thus, for psychological egoism to be true, even actions which advance the welfare of another person, i.e. altruistic actions, must be motivated primarily or solely by perceiving the other person's welfare as having value for advancing the welfare of the actor in some way – extrinsic, proprietary value. If any altruistic action is motivated to any significant degree by the actor perceiving the other person's welfare as having value in itself, without regard to its contribution to anything the actor values – intrinsic, non-proprietary value – then the best explanation for that action is not that the actor is motivated by self-interest, so the universal claim of psychological egoism is refuted by counterexample.

There are many such counterexamples – altruistic actions that are not only motivated by the altruist perceiving the other person's welfare as having intrinsic, non-proprietary value (value it has without regard to its contribution to anything else the altruist values), but that seem to have no other apparent motive. However, the attempt to find "perfectly" altruistic actions with no self-interested component in order to most strongly rebut the psychological egoism thesis sets up a misleading line of counterargument: Defenders of psychological egoism argue that the alleged altruist may have hidden self-interested motives which are not apparent. However, as noted above, any action where the altruist is motivated to some significant degree by perceiving the other's welfare as being of intrinsic, non-proprietary value is sufficient to serve as a counter-example to psychological egoism's universal claim. The refutation of psychological egoism does not require entirely ruling out the possibility that the altruist is also motivated in some part by perceiving extrinsic, proprietary value in the welfare of another, only that there is a significant degree of altruistic motivation – i.e. motivation by perceiving intrinsic, non-proprietary value in the welfare of another.

Similarly, I think rejecting the evolutionary selfishness thesis does not require entirely ruling out the possibility that altruistic or cooperative behaviors also advance the proprietary reproductive *telos* of the altruist. It is sufficient that there exist some behaviors in which an individual in a population of organisms advances the reproductive *telos* of other individuals in the same population. But to see this, again it is worth restating the claim carefully and spelling out the analogous argument against it.

The evolutionary selfishness thesis is, in essence, no more than a rejection of the claim that the reproductive *telos* can ever be other than strictly proprietary. In this, it is much like psychological egoism, which is the rejection of the position that humans can ever act from anything other than self-interested motives. For the evolutionary selfishness thesis to be true, even behaviors which advance another organism's reproductive *telos* at the apparent expense of an organism's own reproductive *telos*, i.e. altruistic behaviors, must ultimately be of extrinsic value with respect to that organism's strictly proprietary reproductive *telos*. Like psychological egoism, this is a universal claim vulnerable to rejection by counterexample.

To understand the argument refuting the evolutionary selfishness thesis, it is important to understand that the reproductive *telos* is not only a value claim, but that it determines everything else of value to an organism. If an organism's *telos* is its own reproduction, that *telos* is the intrinsic value which generates a nested hierarchy of those states of affairs (finding food, avoiding predation, raising young, etc.) which are of extrinsic value to that particular organism insofar as they contribute to its own reproduction. By definition, however, altruistic behaviors are those that do not contribute to an organism's own reproduction, but rather sacrifice that organism's own reproduction to advance the reproductive interest of another organism or organisms within the same population. Thus, an altruistic behavior cannot conceivably be of extrinsic value to that particular organism insofar as it contributes to *its own reproduction* – because insofar as it is altruistic, the behavior detracts from its own reproduction and contributes to the reproduction of another.

The only way sense can be made of altruistic behavior is to understand the reproductive *telos* of the altruistic behavior's beneficiary as being of intrinsic, non-proprietary value to the altruist: Thus, there must be some respect in which an organism's *telos* – that which is of intrinsic value to it and determines everything else of extrinsic value to it – is not simply *its own* reproduction, but rather its own reproduction plus the reproduction of the beneficiaries of its altruistic behavior. Any solution to the problem of altruism which provides an explanation for sacrifice of an individual's fitness to advance the fitness of other organisms – whether those other organisms are its kin or some broader social group – also expands the scope of what is of intrinsic value to it, making that value non-proprietary within that scope (if not entirely non-proprietary).

It is worth noting that, although the strong evolutionary skepticism thesis is refuted, this account does not rule out some essentially selfish component in the way other organisms are valued – no more than the argument against psychological egoism rules out the possibility of a self-interested component in the motivations of even the most ideally altruistic action. Positing that altruistic behavior requires that the reproductive *telos* of other organisms be valuable in an intrinsic, not-exclusively-proprietary fashion does not exclude the reproductive *telos* of other organisms *also* having an extrinsic, proprietary value for the altruist. This, I think, should be sufficient to satisfy the skeptical concern that motivates questions about whether evolved altruistic behavior is "genuinely" altruistic, because it acknowledges that there is still a reproductively self-interested component to the behavior.

The skepticism that motivates the evolutionary selfishness thesis was first raised against the example I used at the beginning of this section, so I will return to that example with the answer: Has my imagined generous uncle simply advanced his own good as defined by his proprietary reproductive *telos*? While he has advanced his own proprietary reproductive *telos*, and so my reproductive *telos* has extrinsic value for him, it is not correct to say he has done "simply" that: The fulfillment of my reproductive *telos* is also of intrinsic, non-proprietary value to him – or rather, less proprietary value, since it still matters that it is of value *to him*; the value of my reproduction *telos* depends on the relationship between my uncle and myself, rather than being of value without respect to any particular valuer.

In effect, expansion of the scope of the reproductive *telos* beyond the individual – as the mechanism of kin selection accomplishes for related individuals – defines a **moral universe**, by which I mean the following: Insofar as the proprietary reproductive *telos* constitutes a normative value claim which can be used as a basis for reaching prescriptive conclusions, it determines what an organism *ought to do* in its own best interest. When the reproductive *telos* encompasses more than just an individual organism, the individuals so encompassed constitute a community of common interest within the bounds of which the value of the reproductive *telos* is non-proprietary. And since a *telos* generates a nested hierarchy of states of affairs which are of extrinsic value with respect to that *telos*, to incorporate the *telos* of other organisms in a non-proprietary fashion is to value *whatever is of value to those organisms as well*. Such an expanded *telos* determines what an organism ought to do not just in its own best interest, but also in the interest of other organisms within the scope of that non-proprietary *telos*.

Thus, while I have used the phrase "less proprietary" a few times above, I did not mean to imply that there is a sliding scale between proprietary and non-proprietary value. Rather, something becomes valuable in a less proprietary way when there is a community of common interest within which it has non-proprietary value; the larger such a community of common interest is, the less proprietary the value which comprises that common interest.

The reproductive *telos* of an organism within a community of common interest is exactly the sort of intrinsic, objective, universal, non-proprietary value claim that can bridge the fact-value gap and serve as the foundation for an ethical theory – hence my decision to call such a community of common interest a moral universe. Such a value claim can be used as a basis for reaching prescriptive conclusions about what organisms within that community *ought to do* in the best interest of its community.

A moral universe is not only a limited-scale model for the foundations of ethical theory, it also reflects the dynamic equilibrium between self-interested pragmatic/instrumental value and other-interested moral value that one would expect from an ethical theory rooted in an even remotely plausible theory of human nature. The moral universe concept does not deny the proprietary nature of the reproductive *telos*, but rather recognizes that the reproductive *telos* need not be exclusively proprietary – that evolutionary mechanisms can create a reproductive *telos* that is also non-proprietary, at least with respect to the community of organisms defined by those mechanisms.

Kin selection alone can account for only very small moral universes, networks of familial obligation that are no less obligatory for their limited scope. Evolutionary mechanisms such as reciprocal altruism and group selection, which I have mentioned but not yet described, can expand the non-proprietary scope of the reproductive *telos* beyond close relatives, accounting for larger moral universes. But, as I shall explain in my next section, these mechanisms alone are insufficient to generate a moral universe which encompasses all of humanity.

Section 3: Natural selection and the limits on expanded moral universes

In his 1971 paper "The Evolution of Reciprocal Altruism," Robert L. Trivers proposed a hypothesis for how natural selection could account for altruistic behaviors that benefit non-kin organisms – even those of a different species – that was quickly and widely embraced by the evolutionary biology community. I will offer only a brief sketch of reciprocal altruism, followed by a slightly more in-depth discussion of group selection theory – just enough to explain why both are inadequate for my ultimate goal of expanding the non-proprietary component of the human reproductive *telos* to a moral universe that includes all humanity.

The key elements necessary for selective forces to favor altruistic behavior towards even non-kin organisms, according to Trivers, are roughly as follows: The definition of altruism at stake is that organisms with an altruistic trait act in a way that on average reduces their inclusive fitness (fitness cost) to increase the inclusive fitness (fitness benefit) of other another organism.

(Requiring a sacrifice of inclusive fitness means that this definition of altruism excludes sacrifices for kin, i.e. behaviors favored by kin selection.) Presuming a population comprised of organisms with and without the altruistic trait, organisms with the altruistic trait do not just incur the fitness cost of their altruistic behavior, but also enjoy the fitness benefit of altruistic behavior from their altruistic compatriots. Selection pressure will favor the spread of such an altruistic trait when the fitness cost and benefits are uneven, such that the altruistic behavior in question incurs a fitness cost to the altruist for each act which is less than the fitness benefit that accrues for each recipient of the act.

Such situations would seem mathematically and biologically unlikely on the face of it: Consider, for example, food sharing. For the purposes of illustration, nutrition can be taken to be numerically equivalent to fitness, such that whatever amount of food one gives up is the fitness cost and whatever amount of food one receives is the fitness benefit. In a given group of organisms where those who have the altruistic trait are presumed to always share with their fellows when the opportunity presents itself and those who lack the trait never share, it is clear that the altruists will always suffer a fitness penalty in comparison to the non-altruists.

But that need not always be the case. If the organisms in question have the capacity to respond to a given non-altruist's failure to share by no longer sharing food with that non-altruist, the benefit-to-cost ratio will change dramatically for both altruists and non-altruists over time: Assuming some randomness in foraging success such that sometimes organisms can benefit significantly from being the recipient of another's food-sharing, down-on-their-foraging-luck altruists are much more likely to receive needed food from their fellows than similarly unfortunate non-altruists. For this to occur, we need not assume any great cognitive capacity in the organisms – just enough to identify individuals, and to remember which ones hoard and which ones share.

Aside from the minimal capacity to remember who does and doesn't reciprocate, a few more key factors are relevant to improving the odds that selection pressure will favor such an altruistic trait in the population. According to Trivers, the benefits of such an altruistic trait are greatest when situations in which the altruist is in a position to benefit another organism occur in the following circumstances:

> ... (1) when there are many such altruistic situations in the lifetime of the altruists, (2) when a given altruist repeatedly interacts with the same small set of individuals, and (3) when pairs of altruists are exposed "symmetrically" to altruistic situations, that is, in such a way that the two are able to render roughly equivalent benefits to each other at roughly equivalent costs. [Trivers, p.37]

As an illustration of circumstances strongly favoring altruistic behavior through the mechanism of reciprocal altruism, consider food sharing behavior in the common vampire bat, *Desmodus rotundus*, as described by Gerald S. Wilkinson in a paper aptly titled "Food Sharing in Vampire Bats." Vampire bats feed exclusively on blood, which is a high-energy food source – but they also have a very high metabolism, and their foraging success is somewhat sporadic (with feeding failure rates on a given night measured at 7% for adults and shown to be highly random for any given individual). For vampire bats inclined to share the results of their foraging with their less successful roost mates (the altruistic trait) – by regurgitating a portion of their blood meal for another bat, pleasantly enough – there are many occasions where this altruistic situation arises in a given bat's lifetime. Because foraging success is largely random, a given altruist is as likely to be an occasionally unsuccessful forager who needs to receive a blood meal as to be a successful forager who has the capacity to share a blood meal. And because the bats have such high metabolic rates and starve to death in 2-3 days without a blood meal, and because the weight loss from fasting increases exponentially as the meager reserves of their tiny bodies are expended, the fitness cost for sharing blood is significantly less than the fitness benefit of receiving blood. Thus, all three criteria mentioned by Trivers are satisfied – and Wilkinson's field observations reveal that this food-sharing occurs extensively even among bats who are not close kin. [Wilkinson, 1990]

While it seems almost incorrect to use the term altruism for this sort of tit-for-tat, "you scratch my back, I'll scratch yours" situation – or rather, this "you vomit up blood for me, I'll vomit up blood for you" situation – it is no less an expansion of the reproductive *telos* to include others in a non-proprietary way than that which occurs via kin selection, and for exactly the same reasons. With that, little more needs be said about reciprocal altruism, and I will move on to group selection.

Imagine a situation that has some significant differences from the vampire bat example: Suppose that the organisms in question don't have the behavioral capacity for recognizing individuals and remembering prior behavior, or that they don't live in the sort of stable groups or engage in the sort of altruistic behavior where reciprocation is a factor. Could they still develop an altruistic behavior towards non-kin? It turns out that they can, but a great deal depends on the organisms' patterns of dispersal.

Consider, for example, a migratory bird of the Northern Hemisphere. Suppose these birds, like snow geese and Canadian geese, gather in great numbers every Summer in remote Arctic breeding grounds that have few predators but ample seasonal food resources. After breeding, they randomly split up into smaller flocks and migrate South to widely dispersed feeding grounds for the rest of the year. While the flocks do include parents and their offspring, the sets of parents and offspring are unlikely to be related to the other sets in the small flocks to a degree greater than chance. Thus, the birds live in largely non-kin groups that are isolated from each other for three quarters of the year or more, but re-gather yearly and shuffle into new groups. (It doesn't matter for the example whether the breeding pairs are stable or themselves reshuffle every year – what matters is the group re-shuffling.)

Suppose these birds exhibit an altruistic behavior of some sort – for example, giving warning calls when they spot predators. Warning calls may slightly increase the caller's chance of drawing the attention of the predator they warn against, but of course having a warning call increases the chances that the rest of the flock will elude the predator – hence the habit of issuing warning calls counts as a fitness-sacrificing altruistic trait.[31] However, it is implausible that warning cries are amenable to any kind of tit-for-tat reciprocal altruism because the altruistic situation does not offer exchange between individuals which can be reciprocated. While there may be many opportunities for altruistic action in an individual's lifetime, the other two criteria cited above do not obtain due to the yearly flock dispersal and shuffling. Also, since the overwhelming majority of the individuals enjoying fitness benefits from the altruistic action are not offspring or otherwise related at greater-than-chance levels, kin selection is unlikely to be a significant selective factor. Yet, selection for an altruistic behavior can still occur under these quite plausible conditions.

To see how selection for this altruistic warning call behavior can occur, let's consider two flocks, one with a low proportion of callers/altruists and another with a high proportion: If the population of birds as a whole has both types in somewhat equal proportions and reshuffles

[31] Some biologists have disputed whether the warning call behavior in birds actually does sacrifice the caller's fitness. Perhaps the individual that spots the predator and thus knows *where* the danger is coming from enjoys a better chance of eluding the danger than its randomly fleeing flock-mates, or perhaps the sheer distraction of the entire flock taking off gives the caller cover from the predator. [Charnov & Krebs, 1975] I think these alternatives are not terribly plausible for larger birds that take off comparatively slowly like geese, but that isn't particularly important: For the sake of my illustrative example, any altruistic behavior whatsoever will do. It is the dispersal pattern of the organism that matters most.

randomly every year, these distributions will represent one flock from each end of the random normal distribution (a bell curve). Suppose each flock has 125 birds – flock **A** having 100 altruists and 25 non-altruists (4 out of 5 are altruists), flock **S** (for selfish) having the reverse numbers (1 out of 5 are altruists). Suppose both flocks suffer 20% attrition from random irrelevant factors such as diseases and jet engines, leaving the proportions of the two flocks at 80/20 and 20/80 respectively, but that all other attrition is due to predation and that warning calls reduce attrition from predation. With fewer callers, largely non-altruist flock **S** will suffer more attrition – but in both flocks, the altruists/callers will suffer proportionally more predation than the non-altruists, and so their numbers will fall as a proportion within each flock.

Let's assign some numbers to reflect these factors: The fraction of altruists/callers in flock **A** is .80, the fraction of altruists/callers in flock **S** is .20. Suppose that the base rate of attrition by predation **PS** is .50 (50% of the flock lost to predators over the course of the non-breeding season) on average without warning calls, but that the chance of a given predator succeeding is significantly reduced by a warning call being issued, such that if warning calls were always issued the predation rate **PA** would be .90 (10% attrition). Assuming that which member of the flock spots a predator first is random[32], then the chance of that bird being an altruist/caller is

[32] Since a predator that is so swift, stealthy or lucky on a particular occasion that it is not spotted at all (by a caller or non-caller) before it strikes is not a factor with respect to warning calls, such predation events can be treated as part of the general attrition rate and ignored. One way to look at this simplifying assumption is that every predator which matters for the purposes of selection for warning call behavior is one that is spotted by some bird in the flock, and the only factor that matters is whether the spotting bird is an altruist/caller or a non-altruist. A more realistic description of the warning-call trait would include recognition that altruists/callers spend more of their time keeping watch and less of their time beak-to-the-ground foraging than non-altruists, potentially representing another fitness cost. Incorporating that more plausible watching-and-calling behavior into the model can be best accounted for by simply taking it to be part of the reason for the large (.50 to .90) attrition rate difference between flocks of all non-altruists and all altruists. If anything, watching-and-calling behavior vs. just calling behavior would seem to make the relationship between the fraction of altruists and attrition rate reduction more positively curved rather than linear – that is, a slight increase in the number of altruists has a larger impact on reducing predation. But my intent is to demonstrate the effectiveness of group selection, so making a simplifying assumption which works against the selection pressure for altruism – yet still shows that altruism is selected for in the end – only strengthens my case.

based on the proportion of altruist/callers in the flock. Thus, the average predation rate over a season for a flock will be based on the projected difference in predation between flocks with no warning callers and all warning callers multiplied by the actual proportion of callers in a given flock. For our two flocks, the average predation over a season will be **PS** + [(**PA** – **PS**) x the fraction of callers in each flock], or .82 for flock **A** and .58 for flock **S**, representing a loss of 18 and 42 birds from each flock respectively.

However, if a predator succeeds despite the warning call, there is at least a slightly greater chance that the individual the predator kills is the altruist/caller – presumably not a vastly greater chance such that the warning caller almost always gets eaten, but some significant risk must be involved for the behavior to be considered altruistic at all. Let's assume that a successful predation event when there is a caller is five times as likely to kill the caller as kill another random member of the flock. That sounds pretty risky, but since it depends on the number of members in the flock, it still isn't a horrible risk: In a 100-bird flock, the caller's chances of dying in a given predation event go from 1 in 100 to roughly 5 in 100 – that is, from .01 to .05, a measly .04 increase. This means that in an evenly split caller/non-caller flock of 100 birds, the odds of a caller dying in a particular predatory event that involves a warning call go from 50/50 to (roughly) 54/46. Of course, those odds are dependent on flock size and will get somewhat worse as predation shrinks the flock – but also remember that those odds only effect predation events where warning cries are issued and so will make less of a difference in flock **S** where there are fewer warning cries. For the sake of simplicity, I will assume that for both groups, the attrition rate is tilted 60/40 such that 3 callers are lost for every 2 non-callers, multiplied by their proportion in the population. Since there are 4 times as many callers as non-callers in **A**, that reflects 12 callers dead for every 2 non-callers. Since there are 4 times as many non-callers as callers in **S**, that reflects 8 non-callers dead for every 3 callers. Please note that, in mathematical terms, this reflects an unrealistically high risk for issuing warning calls and actually tilts the math strongly against the effect of group selection I am about to reveal.

Based on these calculations and approximations, and rounding off fractional birds, the survivors of the two flocks that return to the breeding grounds after their dispersal for the rest of the year would be as follows: The membership of flock **A** contains 82 birds, 65 altruists and 17 non-altruists – the proportion of altruists having dropped from .80 to .79. The membership of flock **S** contains 58 birds, 49 non-altruists and 9 altruists – the proportion of altruists having dropped from .20 to .16. Thus, when these two flocks re-gather at their breeding grounds in the Summer, in total there will be 74 altruists and 66 non-altruists returning from these two flocks to breed the next generation. So even though the fraction of altruists in each group decreased as a proportion of the membership of that group, the fraction of altruists in this sub-population as a whole actually increased from .50 (100 out of 200) to .53 (74 out of 140) due to the fitness benefits of the altruistic behavior as compared between the groups. And since these birds will redistribute more-or-less randomly into new flocks at the end of the breeding season, those new flocks will have a slightly greater average proportion of altruists than they did the year before. The fraction of birds which engage in this altruistic behavior will continue to grow until the point where the between-group comparative fitness benefit of the behavior no longer outweighs its within-group fitness cost.

This hypothetical example, the content of which seems pretty plausible in biological terms and the mathematical details of which were approximated only in ways that hurt the case, demonstrates the mechanics of group selection in action. Generally speaking, group selection for a particular altruistic trait occurs when the benefits of that trait to the fitness of the group as a whole *in comparison with other groups* outweigh the fitness costs to individuals within each group: This can only occur in circumstances where individuals live in reasonably isolated groups, but where individuals move from one group to the next with sufficient regularity – in other words, where there is a population of dispersed groups with some flow of individuals between groups. Since many migratory birds exhibit dispersal patterns very much along the lines I described in my hypothetical example, such dispersal patterns are clearly not impossible – or even particularly unusual.

After offering an illustrative example of group selection for altruistic behavior rather more abstract and formalized than the one I have described above, David Sloan Wilson and Elliott Sober outline the factors necessary for group selection as follows:

> ...What is required to produce this interesting (and for many people counterintuitive) result? First, there must be more than one group; there must be a *population of groups*. Second, the groups must *vary* in their proportion of altruistic types. Third, there must be a direct relationship between the proportion of altruists in the group and the group's output; groups with altruists must be *more fit* (produce more individual offspring) than groups without altruists. Fourth, although the groups are isolated from each other by definition.., there must also be a sense in which they are *not* isolated (the progeny of both groups must mix or otherwise compete in the formation of new groups). These are the necessary conditions for altruism to evolve in the multigroup model. To be sufficient, the differential fitness of groups (the force favoring the altruists) must be strong enough to counter the differential fitness of individuals within groups (the force favoring the selfish type).
>
> These conditions are similar to the ones laid down in standard formulations of Darwin's theory of natural selection, which requires *a population of individuals* that *vary* in heritable characteristics, with some variants *more fit* than others. The analogy extends to the fourth condition, since individuals are isolated units but nevertheless compete in the creation of new individuals. Thus, natural selection can operate at more than one level of the biological hierarchy... Individual selection favors traits that maximize relative fitness within single groups. Group selection favors traits that maximizes the relative fitness of groups. Altruism is maladaptive with respect to individual selection but adaptive with respect to group selection. Altruism can evolve if the process of group selection is sufficiently strong. [Sober & Wilson, *Unto Others*, p.26-27]

I have mentioned that group selection is controversial, but the theoretical and mathematical underpinnings of the selection mechanism outlined above – first convincingly and rigorously formalized by Wilson in his 1975 paper "A Theory of Group Selection" [Wilson, 1975] – are not what is disputed for the most part. Rather, many biologists who acknowledge that group selection is theoretically possible simply deny that it is actual – that is, they doubt that the circumstances under which group selection occurs obtain often enough in the wild for group selection to be an important factor in evolution. However, part of what causes such skepticism is failure to take groups into account at all.

Consider the mechanics of group selection for a trait that does not have a fitness cost – that is, a trait that is not altruistic. A trait which increases the fraction of individuals who possess that trait within a group because it increases their individual fitness will also increase the collective fitness of that group. However, because there is no fitness cost balancing that fitness benefit, one would see the exact same fitness benefit if one simply totaled the fitness of the individuals without regard to their membership in groups; no information would be lost by such a calculation.

In contrast, if one totaled up the offspring of my hypothetical migratory birds without regard to their group membership, one would see the aggregate result and determine that the warning-call behavior has a fitness benefit, but would not recognize that the aggregate result is in fact an aggregate – that the group fitness benefit outweighs the individual fitness cost. On a fitness calculation that simply averaged fitness across groups, the individual fitness cost (the shrinking fraction of altruists within each group) would be entirely invisible – and the fact that the behavior is genuinely altruistic would be likewise invisible, as would the effect of group selection. Similarly, a trait which enhanced the fitness of a group relative to other groups but had no impact on the relative fitness of individuals relative to other members of their group would be also be invisible to someone averaging fitness across groups, and the trait would be mistakenly seen as an individual fitness advantage rather than a group fitness advantage.

In *Unto Others* (quoted above), Sober and Wilson name this error "the averaging fallacy" and relate the history of misunderstanding and dismissal of group selection by those who commit this error repeatedly. [Sober & Wilson, p.31-92] Once they have done so, there is only one element missing from what they name "A Unified Evolutionary Theory of Social Behavior" in the title of their second chapter:

> We have shown that all of the major theories proposed as alternatives to group selection – inclusive fitness theory [a.k.a. kin selection], evolutionary game theory [of which reciprocal altruism is an example], and selfish gene theory [not discussed by me in this dissertation] – merely look at evolution in group-structured populations from different perspectives. In order to combine them in a single unified theory, however, we need a clear definition of groups...
>
> ... In all cases, a group is defined as a set of individuals that influence each other's fitness with respect to a certain trait but not the fitness of those outside the group. Mathematically, the groups are represented by a frequency of a certain trait, and fitnesses are a function of this frequency. Any group that satisfies this criterion qualifies as a group in multilevel selection theory, regardless of how long it lasts or the specific manner in which groups compete with other groups. [Sober & Wilson, p.92-93; my explanatory additions in brackets]

To distinguish this specific understanding of what a group is for the purposes of group selection from other potentially confusing or ambiguous definitions, Wilson coined the term "trait group," which I will use hence. [Wilson, 1975]

By exactly the same reasoning I laid out in the prior section for kin selection, every trait group conferring a selective advantage to some trait with respect to that group defines a moral universe: Whenever a fitness-sacrificing trait of an individual makes a sufficient positive contribution to the fitness of non-related others within a trait group for group selection to occur, the reproductive *telos* of those non-related others is valuable in a non-proprietary way within the bounds of the trait group.

However, for group selection to occur at all, trait groups cannot remain isolated: If there is not some redistribution of birds with the altruistic trait of issuing warning calls at the sight of predators in my hypothetical example – that is, if the flocks did not gather and mix every season in a common breeding ground, but instead just bred amongst themselves year after year – then the fitness cost of giving warning cries would lead to that behavior being represented in smaller numbers of offspring every generation until it disappeared.[33] Thus, in situations where group selection favors an altruistic trait, altruists must carry that trait with them to any new group (or pass it on to offspring who mix with others in new groups). With respect to the altruistic trait itself, whatever group the individual happens to be in is its trait group: A bird that issues warning cries when it sees predators will do so without regard to which particular other birds it is currently flocking with. Doesn't that make the value the altruist bird places on other birds, by virtue of the sacrifice it is making for them, universal and non-proprietary with respect to its whole species?

Or, as another possible expansion of the reproductive *telos*, might an individual organism belong to different trait groups with respect to different traits? That would seem to have some potential to expand the organism's moral universe dramatically, magnifying the number of others in that organism's collective moral universe (or overlapping moral universes, if you prefer) with respect to a constellation of traits.

[33] One might be tempted to think that, with no dispersal of offspring into other groups, the group members will become more related to one another over time and kin selection might start to favor the altruistic trait. But even assuming that the trait is neatly tied to a dominant allele, this doesn't seem likely: Since the fitness cost to the altruists results in their fraction of the population shrinking with every generation, and since the warning call behavior benefits the whole group, the benefit to the altruists' kin will actually shrink with every generation.

Neither of these potential expansions of a moral universe carries any real moral weight, however. Whether an individual organism belongs to different trait groups successively or concurrently, every mode of selection – even group selection – necessarily implies competition and so blocks the expansion of the reproductive *telos* beyond the bounds of the current trait group or groups, quite without regard to how the actual behavior works in practice.

Why is competition necessary? Group selection not only requires distinct trait groups within a broader population of groups, it only operates as a mechanism of selection by virtue of the fitness competition between those groups. Even if a group-selected altruistic trait were to become universally instantiated in a given species, that altruistic trait can only persist in virtue of its continued contribution to group reproductive success *in relation to other groups*. Under different circumstances (environmental change, etc.) where the individual fitness-sacrificing trait no longer benefited group fitness in comparison to groups where the trait is less prevalent, any drift away from that altruistic trait would increase individual fitness without any group fitness penalty and so the altruistic trait would grow less prevalent over successive generations. The evolution of cooperative behavior can only occur when cooperation gives a cooperative group a fitness advantage in comparison to non-cooperative or less-cooperative groups; and the ever-present proprietary core of the reproductive *telos* guarantees that non-cooperative behaviors will always return to dominance if the group-level fitness advantage ever wanes.

This last point is worth a bit of expansion, and it brings up an aspect of my account that has of yet received only indirect attention – disvalue. I have already mentioned in passing the dual nature of the reproductive *telos*: While the reproductive *telos* of other organisms is of intrinsic, non-proprietary value within a given moral universe, it can at the same time be of extrinsic, proprietary value. However, there is necessarily another component of the reproductive *telos* of other organisms that is of extrinsic, proprietary *disvalue* to the altruist, insofar as altruistic behavior is fitness-sacrificing. Altruistic behaviors evolve only when the extrinsic, proprietary disvalue is balanced by a greater extrinsic, proprietary value: Thus anything which undermines the value of the reproductive *telos* of other organisms – that is, anything which undermines the selective pressure in favor of cooperative behavior – necessarily shifts the balance in favor of the disvalue, and therefore undermines the foundation of the intrinsic, non-proprietary value of others within the scope of that moral universe. Group selection does not negate individual competition, it overwhelms it by virtue of greater selective pressure for cooperation within a group based on the advantage it gives that group in fitness competition between groups. When it does so, it expands the proprietary reproductive *telos* of each individual within such a group to include the reproductive *telos* of every other member of the trait group.

So, to the extent that trait groups define moral universes, they only do so in virtue of the fact that each excludes other organisms from that moral universe. We could even suppose that every altruistic behavior failed to differentiate between beneficiaries based on group membership. That is, perhaps every altruistic behavior is like that of my hypothetical altruistic geese, which issue the same warning honks when they spot a predator from one year to the next even though they are surrounded by entirely different geese than the year before – and possibly would still honk even if surrounded by ducks, or starlings, or deer, or no other animals at all. But since the reproductive *telos* is expanded beyond the individual by virtue of mechanisms which depend on competition between trait groups, indiscriminate altruistic behavior towards organisms outside the trait group does not actually confer value on them any more than the indiscriminate imprinting behavior of ducklings confers motherhood on the farmer they follow around.

To bring this discussion back to bear on humanity in particular, which after all is my ultimate goal, this distinction is crucial: Suppose that group selection mechanisms do explain how humans came to acquire certain behaviors or inclinations to treat others altruistically, which is ultimately what Sober and Wilson argue in *Unto Others*, which is subtitled "The Evolution and Psychology of Unselfish Behavior." Suppose further, quite counterfactually, that these altruistic behaviors were directed towards other humans without regard to whether those benefited were affiliated with the altruist in any way that would count as a trait group.[34] Even such a universally altruistic behavior would be just that – behavior. A group selection explanation of that altruistic behavior would not justify any conclusion about what is actually of value to us, it would just be an account of how we acquired the behavior which causes us to treat others in a way consistent with their having value. My argument about objective value does not and cannot depend on or be reduced to beliefs, attitudes or behaviors about values; rather, it is meant to be an account of what is of actual value – at least, if any account of value can be genuinely justified (see the discussion of the conditional nature of my overall argument on p.31).

Thus even if group selection is the correct account of altruistic/cooperative behavior in humans, it cannot possibly result in a genuinely universal, non-proprietary reproductive *telos* because the mechanisms of group selection absolutely require in-group/out-group division in order to function. Similarly, if reciprocal altruism is sufficient to account for human social behavior without any need to appeal to group selection, as Trivers argues [Trivers, p.47], reciprocal altruism also relies on cooperative behavior giving cooperators – or more specifically, reciprocators – a fitness advantage over other organisms, and therefore requires competition to

[34] Of course, if humans really did behave altruistically towards other humans without regard to their affiliations – race, creed, ethnicity, rationality, gender, sexual orientation, political beliefs, and so on – we would hardly have the same drive to understand and develop convincing prescriptive ethical and political theories.

function. Whether expanded by the mechanisms of kin selection, reciprocal altruism, group selection, or all three, natural selection is not best characterized by Hobbes' phrase "the war of all against all" – but it must still involve the war of some against others for selection to occur. An all-encompassing moral universe is impossible to realize by any expansion of the reproductive *telos* through the mechanisms of natural selection operating at any level.

I say "at any level" advisedly, because natural selection does seem to operate at different levels, often with the same trait having a positive fitness effect at one level and a negative fitness effect at another level. In their development of a unified theory of social behavior cited above, Sober and Wilson call this "multilevel selection theory" (see p.123). By this phrase, they mean nothing more (or less!) than a theory which recognizes the different levels at which the same basic components of natural selection – variation which leads to differential fitness within a population over generations – can operate. Selection can operate on genes, as described in selfish gene theory (which I have not touched on here); on individual organisms through the ordinary processes of natural selection; on individual organisms and their relatives through kin selection; on individual organisms which interact repeatedly with other organisms through reciprocal altruism; and on groups which disperse and mingle through group selection. Stephen Jay Gould has even argued –persuasively, in my opinion – that a form of selection also operates on species, and perhaps even at higher taxonomic levels. [Gould, 2002] However, I do not think that analyzing species-level selection mechanisms would be an effective or particularly convincing way of getting at a genuinely non-proprietary value claim foundation for ethical theory. Rather, having explored how natural selection operates to expand the reproductive *telos* somewhat, I believe that further expansion can only be found where the operations of natural selection break down.

Section 4: Cultural evolution and universal morality

In *Darwin's Cathedral*, David Sloan Wilson makes an extended argument that a very complex collective human behavior – religion – can be analyzed as the result of group selection operating on human culture. [Wilson, 2002] I am not particularly interested in the argument about religion as such, but the idea of cultural selection has implications which I think are very important for my project, and which Wilson does not quite seem to recognize.

One thing Wilson *does* recognize is that group selection is not a proper foundation for ethical theory:

> [E]ven when groups do evolve into adaptive units, often they are adapted to behave aggressively towards other groups... Group selection does not eliminate conflict but rather elevates it up the biological hierarchy, from among individuals within groups to among groups within a larger population. The most that group selection can do is produce groups that are like organisms in the harmony and coordination of their parts. We already know about the competitive and predatory interactions that take place among individual organisms in ecological communities, and the same can be expected of well-adapted groups. This might be a disappointment for those searching for a universal morality that transcends group boundaries, but it follows directly from the organismic concept of groups. I do not mean to imply that the search for a universal morality is hopeless, only that it does not follow automatically from group selection theory. [Wilson 2002, p.10]

Indeed it does not.

The basic idea of what Wilson here calls "the organismic concept of groups" was already implicit in my discussion of group selection in the prior section: Groups are, like individual organisms, units that can vary, and those variations can have fitness consequences with respect to other units in the population. However, simply undergoing group selection at all is not enough to "produce groups that are like organisms in the harmony and coordination of their parts." As Wilson says, that is the *most* that group selection can do. But how can group selection do even that much? The story of group selection I told in the prior section was characterized primarily by a tug-of-war between opposing selective forces – an altruistic trait has a group fitness benefit with respect to other groups sufficient to overcome the relative individual fitness cost with respect to other individuals within a group. It seems intuitive that such a war of opposing selective forces cannot tilt so strongly towards cooperation.

However, that intuition is based on an incomplete understanding of group selection. The same sort of mechanisms that sustain genetic stability in individual organisms – in spite of the opposing selective force on genes to replicate themselves indefinitely using all the resources available – can also operate at the group level to generate group stability:

> For example, a bacterial cell can be regarded as a social group of genes that coordinate their activities for their collective benefit. However, this group can be exploited by genes that use the resources of the cell to replicate themselves rather than by making products that contribute to the common good... [T]his problem can be solved by linking all the genes together into a chromosome that replicates as a unit. By eliminating the possibility of differential replication within the cell, chromosomes concentrate the process of natural selection at the among-cell level, neatly solving the fundamental problem of social life. But the genes responsible for the evolution of the chromosome do not appear self-sacrificial. Instead, they appear to benefit the group, of which they are a part, at no special cost to themselves...
>
> Social control, rather than highly self-sacrificial altruism, appears to solve the fundamental problem of social life at the individual level... What works for individuals can also work for social groups. In their drive to explain highly self-sacrificial altruism, sociobiologists have tended to ignore an even more important question: Does benefiting the group require overt altruism on the part of individuals? If not, then group selection can favor mechanisms that organize groups into adaptive units without strong selection against these mechanisms within groups. [Wilson 2002, p.18- 19]

As I noted in the previous section (see my discussion of the averaging fallacy starting on p.122), the motivation for group selection theory may be to explain the evolution of altruistic traits which sacrifice individual fitness for the benefit of other individuals, but it is possible for a trait to benefit group fitness without having any significant impact on fitness relative to other individuals in the group. What sort of traits can benefit a group at no particular cost to an individual? For one example, let's look at mechanisms of social control such as punishment and shunning.

In reciprocal altruism, the altruist who aids the non-altruist that fails to reciprocate learns not to aid that individual any more. But what if the other altruist members of the group also learned that lesson by some method of observation or communication, and so also refused to aid non-reciprocators, effectively shunning them? One or two instances of receiving aid without reciprocating and a non-reciprocator would never be helped again by any individual who had both the altruist trait and the shunning trait. The non-reciprocator types would benefit substantially less in a group where the altruist trait was paired with a shunning trait – and because the individual fitness cost of altruism must be measured in comparison to other group members (otherwise we commit the averaging fallacy), a lower fitness benefit for non-altruists effectively means a lower fitness cost for altruists. Better still, the individual fitness impact of the shunning trait itself is at least slightly positive, because shunners do not expend their resources to benefit those who have already revealed themselves to be non-reciprocators. Thus, the relative within-group fitness disadvantage of an altruistic trait is reduced by being paired with a shunning trait or some other mechanism of social control. This effectively transforms the reciprocal altruism mechanism of the unspecified altruistic trait into a group selection mechanism, and amplifies the group selection benefits of both the altruistic trait and the shunning trait.

Social control mechanisms are just one kind of trait that can play a powerful role in group selection by creating a fitness benefit for a trait group without a commensurate individual fitness cost, but they play a very important role in Wilson's overall argument in *Darwin's Cathedral*: In humans, the social control mechanisms that create group cohesiveness and result in human communities acting more like an integrated adaptive unit are *moral customs*. Because they represent the oldest form of human culture (and presumably the cultural template for our pre-human ancestors as well), Wilson focuses on hunter-gathererer groups as his first illustrative example:

> So far I have discussed basic evolutionary principles that apply to all organisms. Now it is time to focus on our own species. We evolved in small groups that are roughly approximated by modern hunter-gatherer societies, which, although disappearing fast, still dot the surface of the globe... [Wilson 2002, p.20]

After a few paragraphs discussing the overwhelming anthropological evidence that the default hierarchy within hunter-gatherer societies is none at all, i.e. they are egalitarian, Wilson continues:

> Hunter-gatherers are egalitarian, not because they lack selfish impulses but because selfish impulses are effectively controlled by other members of the group. This form of guarded egalitarianism has been called "reverse dominance" by anthropologist Chris Boehm. In many animal groups, the strongest individuals are usually able to dominate their rivals, taking a disproportionate share of the resources. This is within-group selection pure and simple. In human hunter-gatherer groups, an individual who attempts to dominate others is likely to encounter the combined resistance of the rest of the group. In most cases even the strongest individuals is no match for the collective, so self-serving acts are effectively curtailed.
>
> ...Boehm explains egalitarianism in terms of social norms, a shared understanding or do's and don'ts that are enforced by rewards and punishments. A hunter-gatherer society is above all a moral community with a strong sense of right and wrong that organizes the practices of the group. The specific practices regarded as right and wrong might vary across groups, but in general "right" coincides with group welfare and "wrong" coincides with self-serving acts at the expense of other members of the group. [Wilson 2002, p.21-22]

None of this is particularly controversial, so I will let it stand without further comment. My interest (and Wilson's) lies primarily not in particular moral systems, but in building an understanding of the process of cultural evolution that generates moral systems. Toward that end, the first point Wilson makes about moral customs is simply that we do observe them in all human cultures, so the underlying psychological mechanisms which make them possible must be universally present in humans. He readily admits that the science describing what he calls "the innate psychology of moral systems" is far from well-developed, although many elements of the system have been extensively researched, such as: the inclination to feel sympathy, a sense of fairness, the mental tools to identify cheaters and the inclination to punish them, and so on.[35] [Wilson, p.26] But whatever the details of the innate psychological mechanisms which allow humans to live together in cohesive communities characterized by moral systems, the most important fact about those mechanisms is that they are capable of underlying a great variety of different moral systems. While hunter-gatherers tend to be egalitarian, the specific ways in which egalitarian behavior is encouraged and enforced varies wildly – and, obviously, not all humans live in hunter-gatherer cultures. The existence of some (not completely specified) innate psychology of moral systems is the "first basic fact" alluded to below:

> The second basic fact that we must understand from an evolutionary perspective is that moral systems include an open-ended cultural dimension in addition to an innate psychological dimension. Our genetically evolved minds make it possible to have moral systems, but the specific contents of moral systems can change within groups and vary widely among groups, with important consequences for survival and reproduction. Far from leading to the caricature of genetic determinism that limits the capacity for change, our innate psychology creates a capacity for change by setting in motion a process of cultural evolution. [Wilson, p.28]

Although I will soon express a profound disagreement with Wilson about his characterization of this process of cultural evolution, I certainly do not disagree with him on these basics: The specific contents of moral systems do vary from group to group and have important consequences for survival and reproduction, and this does set the process of cultural evolution in motion. Moreover, I find Wilson's explanation of the open-ended nature of cultural evolution very useful and instructive. He begins with an analogy:

[35] There is no reason to think that these innate psychological mechanisms are limited to humans. In my introduction, I mentioned research showing that something like a sense of fairness, or at least the capacity to recognize and be displeased by manifest unfair treatment, has been experimentally demonstrated in several non-human social animals (see p.5).

> Consider the mammalian immune system. Just like the mind, it can be regarded as a collection of specialized genetically evolved mechanisms for helping us survive and reproduce in our ancestral environment...[T]he centerpiece of the immune system is an open-ended process of blind variation and selective retention. Antibodies are produced at random and those that successfully fight invading disease organisms are selected...
>
> This comparison, between the mind and the immune system, is simple but profound in its implications. It shows that genetic evolution does not invariably lead to the kind of modularity that excludes open-ended processes. Instead, it can create processes that are themselves evolutionary and therefore capable of providing new solutions to new problems. Plotkin has aptly termed these processes "Darwin machines"...
>
> Cultural evolution can be seen in part as a Darwin machine in action, highly managed but nevertheless genuinely open-ended in its outcome. Confront a human group with a novel problem, even one that never existed in the so-called ancestral environment, and its members may well come up with a workable solution. The solution might be based on trial and error or on rational thought. However, rational thought is itself a Darwin machine, rapidly generating and selecting symbolic representations inside the head. Confront many human groups with the same novel problem and they will come up with different solutions, some much better than others. If the groups are isolated from each other, they may never converge on the best solution; evolution is not such a deterministic process. If the groups are in contact, they might compare solutions and the worst might quickly imitate the best. If convergence by imitation does not occur, then the worst might simply succumb to the best in between-group interactions. Either way, the final outcome is a degree of adaptation to the problem without any genetic evolution taking place at all. Evolution took place, but not at the genetic level. [Wilson 2002, p.31-33]

Again, I not only agree with this characterization of the open-ended nature of cultural evolution, but find it very insightful and useful. But throughout the remainder of his argument, Wilson treats cultural evolution as just another kind of group selection – and he seems to have glossed over a major distinction to do so. The issue he fails to acknowledge is exposed by his discussion of imitation above: **Culture transcends inheritance.**

All levels of natural selection, including group selection, operate through the mechanism of offspring *inheriting* the traits which have an impact on fitness. While human culture is transmitted from parents to offspring (rather unreliably in many cases), it is also transmitted between individuals without reproduction; the means of cultural transmission is communication rather than reproduction, and the content transmitted is ideas rather than genes, embodied in symbols rather than DNA. Wilson acknowledges this in passing, but doesn't directly or fully acknowledge that this is a crucial difference that alters the workings of natural selection at every level, and so cannot be ignored or bracketed to treat cultural evolution as a subset of group selection.

I think that Wilson's view of human culture as a collective "Darwin machine" that generates potential adaptive responses to changing circumstances is both a useful analogy and an accurate characterization of the transformative power of culture. However, the results of that process cannot be treated as traits which undergo group selection, because group selection – like all natural selection – depends on traits being passed down to offspring. Cultural traits, in contrast, are passed around as well as passed down: That is not to deny that cultural traits do undergo some sort of selection – certainly they do, and phrases like "cultural evolution" and "cultural selection" seem apt. But ideas flow much more freely than genes, so cultural selection must operate quite differently than natural selection.

Before proceeding further, it might be useful to disambiguate the term 'culture,' which is used in many different senses and has many different definitions even within cultural anthropology, the science most prototypically identified with the study of culture. For my purposes here, I will use the broadest, most inclusive definition of the term prominent in anthropological literature – the definition most compatible with Wilson's deliberately broad use of the term – as articulated by British anthropologist Edward B. Tylor in 1874:

> Culture or Civilization, taken in its wide ethnographic sense, is that complex whole which includes knowledge, belief, art, morals, law, custom, and any other capabilities and habits acquired by man as a member of society. [Tylor, p.1]

While Tylor uses the words "capabilities" and "habits," the common theme that unites the listed examples is that they are all first and foremost *ideas* (in the representational sense): Knowledge and beliefs are self-evidently so. Art may be embodied in a physical work of art or performance, but it must be conceived before it can be executed. Laws and customs may structure institutions and activities, but that structure comes from complex mental representations. Anthropologists also frequently talk about 'material culture' – especially those who specialize in archeology, for obvious reasons – but our tools and infrastructures and so on must be conceived and built, and what humans pass down (or across) is not just the physical objects, but also the knowledge and skills required to make them. Culture consists in both ideas themselves and in the actions and activities which those ideas shape, and thus we can treat ideas as the traits which vary and can be selected (although they are not necessarily inherited).

Defined in this broad sense, culture is unique to humans – but that is a matter of contingent historical fact rather than being a necessary and essential characteristic of humanity. Indeed, some aspects or parts of culture can be seen in other animals: Chimpanzees appear to have local customs or traditions embodied in different techniques used to crack nuts in the wild (and other complex learned behaviors), and experiments have shown how these traditions can be transmitted between groups by demonstration and imitation. [Whiten *et* al, 2007] However, the extraordinarily flexible and free-flowing generation and exchange of ideas allowed by human language – the representational capacity which underlies the broad range of ideas captured by the concept of 'culture' defined above – has to date only been discerned in humans.[36]

[36] There is little dispute that language itself is unique to humans, but the communicative capacities which form the foundations from which language evolved are, unsurprisingly, widely shared by other social organisms. I say "unsurprisingly" because shared or similar characteristics, either inherited from a common ancestor or generated by similar ecological interactions (convergent evolution), is a predictable outcome of the processes of evolution. The nature and evolution of language is hardly a settled area of science, but for one interesting and influential discussion which focuses on the differences between protolinguistic capacities in other organisms and full-blown language in humans, see Derek Bickerton's

Culture – like Wilson's analogous Darwin machine, the antibody-generating part of our immune system – generates alternatives and explores the space of possibilities. Many different cultural traits – different ways of representing the world and attempting to solve the problems it poses – can be generated and put into practice by an individual over the course of a single lifespan or by a group over the course of a generation, and the next generation may abandon many of them for entirely new ones. The only limits on this process seem to be contingent and fleeting: Language barriers fall to migration and trade and new communication technologies, and cultural barriers cannot easily and consistently withstand sustained outside influences.

Cultural selection, like antibody selection, is the process of sorting through the alternatives that have been generated for what is useful. 'Usefulness' is a functional term – an Aristotelian 'for the sake of which' – and for living things, function is determined by the reproductive *telos*. Cultural traits/ideas will undergo positive selection to the extent that they advance or are otherwise consistent with what is of value to individual humans (or human groups which comprise moral universes), and negative selection when they hinder or are otherwise inconsistent with the same. Insofar as the criteria for cultural selection are still rooted in the reproductive *telos*, cultural selection parallels natural selection in some sense. But because the cultural traits selected are not constrained within lineages, the mechanisms of cultural selection are not identical with any level of natural selection. Cultural selection is thus dependent on natural selection, but is not directly derivative of or reducible to natural selection – in other words, cultural selection *supervenes* on natural selection and the proprietary reproductive *telos* natural selection creates.

How exactly do the mechanisms of cultural selection differ from natural selection? With respect to cultural traits, humans are neither individuals nor members of trait groups that can be

Language and Human Behavior [Bickerton, 1995].

reproductively isolated as units of selection, but participants in a somewhat free-form trait-swapping cultural collective. The lineage-crossing nature of cultural traits makes all participants in culture (individually and in groups) potential sources for and recipients of solutions to collective and individual survival and reproduction challenges. Like the antibody-generating components of the immune system, the value of participation in culture lies not in any particular (cultural) trait participants create or possess, but in the diversity and ongoing creation of cultural elements – ideas – which have the potential to solve any problem an ever-changing world can generate.

Here, then, is the basis for formulating a highest human good, the non-proprietary component of the reproductive *telos* which encompasses all humans and so finally fulfills the criteria for a foundational value claim on which an ethical theory might be built: not cooperation within any particular group or culture, but participation in the production and exchange of culture. The characteristic on which the value of every human being to every other human hinges is the same characteristic which most scientists and philosophers in one way or another see as most characteristic of and unique to humanity – not the human mind in its totality, but the part of the mind "that has articulate speech," which Aristotle called the rational soul (see p.53).

Section 5: An evolutionary foundation for ethical theory

At this point, I would like to step back and reiterate the goal towards which I have been working, so I can satisfy myself (and hopefully my readers) that I have achieved it: In Chapter 2, Section 2, I developed a solution to the challenge Hume raises for the justification of any ethical theory. That solution hinged on identifying some state of affairs which has intrinsic, objective value for each and every human as a basic normative premise from which prescriptive conclusions could be justifiably derived, potentially culminating in the universal prescriptive claim(s) of a complete ethical theory. Identification of the normative claim as being true for everyone creates the possibility of arguing towards not just prescriptive claims, but **universal** prescriptive claims which are true for everyone. Identification of the normative claim as **intrinsic** makes it possible to determine the weight of various potential normative claims in order to arrive at those universal prescriptive claim conclusions – as well as reflecting the common position within the tradition of ethics that ethical precepts override prescriptive claims derived from lesser concerns, e.g. other normative claims which are either extrinsic or subjective, possibly both. Finally, identification of the value claim in question as **objective**, i.e. factual, would answer the primary Humean concern that the relation between fact and value "shou'd be observ'd and explain'd." What remained was that "a reason should be given," or that **justification** be offered for some particular foundational value claim or claims. (See p.29 for details and context.) Later, I refined the understanding of universal and objective value to explain how they conjoin in **non-proprietary** value, which proved such a crucial and instructive concept that I decided to treat it as a criterion in its own right.

The state of affairs of intrinsic, objective, universal, non-proprietary value I have identified is participation in culture, or the creation and dissemination of ideas, which comprises the non-proprietary part of the reproductive *telos*. In other words, humanity's shared *telos* is the creation/reproduction of culture, rather than the mere reproduction of individual humans (or of selectively circumscribed human groups). "Universal" here does not simply mean that everyone values their own participation in culture or just participation in their own culture, either of which embody the confused sense of universality which leads only to proprietary value. Rather, participation in culture itself – no matter whose participation and no matter which culture – is of value to all other participants in culture. Participation in culture is of objective value because that value derives from the reproductive *telos*, and that every organism's *telos* is its own reproduction is a matter of fact justified by the same inductive arguments which support the theory of evolution by natural selection, and based upon the same empirical evidence (see p.88-90). The value of participation in culture is intrinsic because it is simply the non-proprietary component of the reproductive *telos*: A being's *telos* is its highest good, that which is of intrinsic value to it, that by virtue of which every other valuable state of affairs has (extrinsic) value for that being.

The conjunction of objective and universal value implies non-proprietary value (see p.42), but it will be especially useful to discuss this criterion separately. The expansion of the proprietary reproductive *telos* to include the reproductive *teloi* of other organisms and so be non-proprietary within a limited scope (a kin group, reciprocal exchange group, or trait group) was constrained within those limits by the operations of natural selection. In removing the connection between descent and the traits which undergo selection, the operations of cultural selection remove the limits on the scope of other organisms which can be encompassed: Cultural traits/ideas which have some fitness benefit, unlike the traits subject to natural selection, can have that same fitness benefit for non-descendants who participate in cultural exchange – even those outside any and every kin group, reciprocal exchange group, or trait group. These limited moral universes of conjoined reproductive *teloi* – even the completely proprietary reproductive *telos* that I might metaphorically refer to as a moral universe of one – are united by cultural selection into a greater moral universe encompassing all participants in the production and exchange of culture. The proprietary reproductive *telos* takes on this universally non-proprietary component because cultural traits can be passed along without reproduction, but the "selection" in cultural selection still depends on the hierarchy of values determined by this collectivized reproductive *telos*. That is, morally valuable cultural traits are those which are good for humanity, not just those good for some particular human or humans.

Until there was some respect in which the reproductive *telos* was fully non-proprietary for all humans (not just non-proprietary within a certain limited scope), the only sense of 'value' determined by that reproductive *telos* was instrumental or pragmatic value rather than moral value – good for some particular organism(s) rather than simply good. However, the non-proprietary value of participation in culture doesn't eliminate the less-proprietary and wholly proprietary components of the value of the reproductive *telos*: The same trait can be valuable in one respect and disvaluable in another, as exemplified in my bird example (see p.116+), in which alarm calls were valuable with respect to the collective reproductive *telos* of each flock as a trait group (compared to other groups) but disvaluable with respect to the reproductive *teloi* of individual group members (compared to other members of the group). So how are these values to be weighed?

Insofar as the justification of any ethical theory is possible – the antecedent of my overall conditional argument (see p.31) – the non-proprietary component of value which expands a being's moral universe beyond the limits of group competition to become genuinely universal *overrides* all proprietary components of value. That is not to say that the operations of cultural selection actually result in cultures where universal non-proprietary components of value overcome proprietary components of value: The overriding character of non-proprietary value is an 'ought' claim, not an 'is' claim – which sets it quite apart from the other levels of selection I have discussed.

The reproductive *telos* is not just the foundation for the foundational fact-value bridging claim that participation in culture is of intrinsic, objective value to every human; it is the basis for a comprehensive theory of value. The reproductive *telos* also determines all other states of affairs that are of objective value to individual humans and human groups – not of objective moral value, but of objective practical value with respect to individual and/or collective self-interest. Those other valuable states of affairs are only practical and not moral because, even though they are of intrinsic and objective value, they are not of universal, non-proprietary value. That does not in any way mean that these other values are inherently immoral, but they are non-moral – and what is of moral value sets limits on what anyone *ought* to do in pursuit of what is of merely practical, self-interested value.

All ethical theories set boundaries on the pursuit of self-interest – but they do so in different ways and for different reasons. Exactly what limits the value of participation in culture sets on the pursuit of self-interest isn't entirely obvious, but one direct implication is that one ought not act in ways that limit or prevent other people from participating in culture. Because culture is such an open-ended activity – generating and acquiring knowledge, inventing and sharing beliefs, art, customs, and so on – and since the value of culture lies in its diversity and open-ended character, a very broad respect for individual autonomy is strongly implied by my foundational value claim, as is an ethical imperative to allow and encourage people to develop their intellectual, artistic, and other potentials to the fullest. These are more broad ethical notions than concrete ethical principles, but it is difficult to say more without going through some kind of step-by-step development of a full ethical theory, which is well beyond the scope of this already large project.

In the absence of a fully-developed ethical theory, I think just one more broad ethical notion is worth discussing. I have so far emphasized what is valuable in my discussion of cultural selection, but it is also worth considering what has negative value. Like the generation of diverse antibodies, the generation and dissemination of diverse ideas serves as a Darwin machine that produces the raw material on which selection can operate at a vastly accelerated rate. That accelerated capacity to adjust and react to changing circumstances makes the activities of the cultural Darwin machine very valuable with respect to the reproductive *telos*, but the individual ideas generated, like the individual antibodies produced by the immune system, might be valuable or neutral – or even disvaluable, as an antibody that causes an autoimmune disease. Thus, while the generation and exchange of ideas – i.e. participation in culture – has great value, that does not mean that every idea generated and exchanged has great value (rather than being value-neutral or disvaluable).

One obvious implication of this reasoning is that cultural traits/ideas which themselves limit participation in the creation and exchange of culture have negative value: For example, sexist or racist ideas which, when put in practice, deny some people education and basic freedoms not only conflict with the universal value of participation in culture because they limit participation for some people, they also directly undermine the diversity that makes culture so valuable in the first place. Putting such ideas into practice also requires actions that violate the broad principles already noted, insofar as they undermine autonomy and actively discourage the development of human potential, so perhaps this self-referential judgment of cultural traits with respect to the value of participation in culture is redundant. However, it's not necessarily a bad sign that the foundation for ethical theory I have developed support widely accepted moral principles in multiple ways.

CHAPTER 5

CONCLUSION

Section 1: Answering the "Open Question"

The goal of this extended argument has been, as I stated from the beginning, to reach an understanding of human nature grounded in evolutionary biology which establishes firmer foundations for ethical theory than any of the traditional contenders. I chose not to address general doubts about the legitimacy or possibility of achieving this goal in advance, simply because the best answer to any claim about the impossibility of some task is to accomplish it. Any remaining questions about whether the task has been accomplished legitimately or fully – or accomplished at all – can be addressed much more clearly once the specifics have been presented.

Perhaps the most prominent of such remaining questions is G.E. Moore's famous "Open Question Argument." In his influential 1903 book *Principia Ethica*, Moore claimed that any attempt to define or explain moral good by reference to anything else is impossible, and dubbed all attempts to do so examples of what he called "the naturalistic fallacy." While Moore's position obviously opposes my argument here, it also denies the foundational value claims of Aristotle, Kant and Mill – all of whom do attempt to explain the nature and origin of the good as the foundation for their claims about the right. Given the sweeping nature of Moore's objections with respect to the entire tradition of ethical theory, I think it would be beneficial to examine some of the context and background of his position before looking at the Open Question Argument proper.

Moore divided the general motivating question of ethics I cited at the beginning of this paper – How should one live one's life? – into two separate but related questions:

> I have tried in this book to distinguish clearly two kinds of question, which moral philosophers have always professed to answer, but which, as I have tried to shew, they have almost always confused both with one another and with other questions. These two questions may be expressed, the first in the form: What kind of things ought to exist for their own sakes? the second in the form: What kind of actions ought we to perform? I have tried to shew exactly what it is that we ask about a thing, when we ask whether it ought to exist for its own sake, is good in itself or has intrinsic value; and exactly what it is that we ask about an action, when we ask whether we ought to do it, whether it is a right action or duty. [Moore, 1929, p.vii]

In other words, Moore is here distinguishing the good from the right, normative claims from prescriptive claims, an account of what is intrinsically valuable/good from an ethical theory about what principle(s) ought to guide our actions. He goes on to say, as one would expect, that the latter are based on the former – that one can only determine which actions ought to be taken by reasoning from established value commitments. I do not disagree with this distinction in the slightest, of course: I make the same distinction in my own argument for the common nature of foundational value claims (the good), and in my analysis showing how representative versions of utilitarianism, deontology and virtue theory all base prescriptive value claim conclusions (the right) on those foundational value claim premises.

Note, however, that my analysis also directly undercuts Moore's claim that other moral philosophers have confused the good and the right with each other or with other issues. In what follows, I will show that Moore's criticism of other moral philosophers springs from confused views on the nature of the good, a confusion which begins to be evident in the paragraph immediately following the quotation cited above:

> But from a clear insight into the nature of these two questions, there appears to me to follow a second most important result: namely, what is the nature of the evidence, by which alone any ethical proposition can be proved or disproved,

> confirmed or rendered doubtful. Once we recognize the exact meaning of the two questions, I think it also becomes plain exactly what kind of reasons are relevant as arguments for or against any particular answer to them. It becomes plain that, for answers to the first question, no relevant evidence whatever can be adduced: from no other truth, except themselves alone, can it be inferred that they are either true or false. [Moore, 1929, p.vii]

For me, this perspective raises two immediate problems. The first is a matter of moral epistemology very close to the heart of this project's concerns: If there can be no evidence or inference – no argument whatsoever – offered in support of a claim about the good, how can any such claim be justified? Moore contends that the good can only be known by intuition. Like many philosophers, I tend to be immediately and deeply suspicious of appeals to intuition in general, and especially to claims that some or other claim of central importance can only be known by intuition. Even though Moore limits his intuitionism to claims about the good rather than extending it to claims about the right (unlike other moral philosophers usually labeled intuitionists), the general argument against appeals to intuition I made in my introduction still applies: Our intuitions about value claims cannot refer to the world independent of our opinions, and thus founder at the slightest disagreement between different people's intuitions – which in turn reveals that even a consensus on intuitions is but an appeal to agreement rather than any kind of substantial justification. Intuition can serve as a useful feedback mechanism and aid to conceptual clarity, but in the end intuition can justify little or nothing – value claims least of all. (See p.11+ for a more complete version of this argument.)

From my perspective, then, Moore falls in the camp of those who reject the possibility justifying any ethical theory, denying the antecedent to my overall conditional argument (see p.31). Moore would almost certainly not see himself as denying the possibility of justifying an ethical theory, since he does take claims about the right to be justified by the truth of claims about the good. But his position that the truth of claims about the good can *only* be known by intuition means that someone who rejects intuition as a way of ascertaining the truth of a claim – and I do – must interpret his position as offering no avenue for the justification of claims about the good or the right.

The second problem in Moore's position is a matter of moral metaphysics also very close to the heart of this work's concerns. What sort of concept is this 'good' that can only be known by intuition, and cannot be explained in any way? Moore insists that there is no possible analysis, there is only the concept of 'good' itself, which is simple and undefinable – which is why it can only be known by intuition:

> ... If I am asked, 'What is good?' my answer is that good is good, and that is the end of the matter. Or if I am asked 'How is good to be defined?' my answer is that it cannot be defined, and that is all I have to say about it. But disappointing as these answers may appear, they are of the very last importance... [I]f I am right, then nobody can foist upon us such an axiom as that 'Pleasure is the only good' or that 'The good is the desired' on the pretence that this is 'the very meaning of the word.'
>
> **7.** Let us, then, consider this position. My point is that 'good' is a simple notion, just as 'yellow' is a simple notion; that, just as you cannot, by any manner of means, explain to anyone who does not already know it, what yellow is, so you cannot explain what good is. Definitions of the kind that I was asking for, definitions which describe the real nature of the object or notion denoted by a word, and which do not merely tell us what the word is used to mean, are only possible when the object or notion in question is something complex. You can give a definition of a horse, because a horse has many different properties and qualities, all of which you can enumerate. But when you have enumerated them all, when you have reduced a horse to his simplest terms, you can no longer define those terms. They are simply something which you think of or perceive, and to anyone who cannot think of or perceive them, you can never, by any definition, make their nature known. [Moore, 1929, p.6-7]

To summarize a bit before continuing, what Moore seems to be saying here is that there is simply no explaining *why* something is good. If someone claims "**X** is good," questions along the lines of "What makes **X** good?" or "Why is **X** good?" are simply nonsense: Asking such a question constitutes a category mistake, because 'good' is not a quality that can be explained or defined.

For just one example of how this position is problematic, consider: How could any claim about right action – that is, a prescriptive claim about what anyone or everyone ought to do in a given circumstance – be grounded in a claim about the good that is not universal? In fact, Moore does believe that "all judgments of intrinsic value are... universal" [Moore, 1929, p.27]. But if that is true, there is certainly a sense in which one of the features that renders a claim about the good legitimate or plausible is its universality, and that is already more analysis than seems permissible under Moore's rigid position that good cannot be analyzed, defined or explained. That is, the concept 'good' must be more analyzable than Moore claims.

A passage from *Principia* a few pages further along seems to offer an answer for this critique – and to some extent, it does. But in doing so it reveals even deeper conceptual problems:

> **10.** 'Good,' then, if we mean by it that quality which we assert to belong to a thing, when we say that the thing is good, is incapable of any definition, in the most important sense of that word. The most important sense of 'definition' is that in which a definition states what are the parts which invariably compose a certain whole; and in this sense 'good' has no definition because it is simple and has no parts. It is one of those innumerable objects of thought which are themselves incapable of definition, because they are the ultimate terms of reference to which whatever is capable of definition must be defined... There is, therefore, no intrinsic difficulty in the contention that 'good' denotes a simple and indefinable quality. There are many other instances of such qualities.
>
> Consider yellow, for example. We may try to define it, by describing its physical equivalent; we may state what kind of light-vibrations must stimulate the normal eye, in order that we may perceive it. But a moment's reflection is sufficient to shew that those light-vibrations are not themselves what we mean by yellow. They are not what we perceive. Indeed, we should never have been able to discover their existence, unless we had first been struck by the patent difference of quality between the different colours. The most we can be entitled to say of those vibrations is that they are what corresponds in space to the yellow which we actually perceive.
>
> Yet a mistake of this simple kind has commonly been made about 'good.' It may be true that all things which are good are also something else, just as it is true that all things which are yellow produce a certain kind of vibration in the light. And it is a fact, that Ethics aims at discovering what are those other properties belonging to all things which are good. But far too many philosophers have

> thought that when they named those other properties they were actually defining good; that these properties, in fact, were simply not 'other,' but absolutely and entirely the same with goodness. This view I propose to call the 'naturalistic fallacy' and of it I shall now endeavour to dispose. [Moore, 1929, p.9-10]

Moore's claim that "Ethics aims at discovering what are those other properties belonging to all things which are good" would seem to answer my previous objection: The claim that 'Universality is a property of all claims about the good' would seem to be a way in which Moore does allow for some conceptual analysis of 'good.' But Moore's denial that those properties in any sense *define* goodness – that they are what makes good things 'good' – would seem to go against such an interpretation. We are back to Moore denying the possibility of defining good – which seems to deny that we can establish any sort of criteria for judging or evaluating proposed claims about the 'good,' or even distinguish them from other claims – while at the same time declaring that the business of ethics is to discover the common features of all good things. One begins to suspect that something has gone seriously awry in Moore's understanding of definition and conceptual analysis. In that opinion, one would have good company – an older G.E. Moore, who in a 1932 paper titled "Is Goodness a Quality?" admitted that his "supposed proofs [that good was indefinable] were certainly fallacious." [Moore, 1959, p.89]

So what exactly is wrong with (the younger) Moore's understanding of 'good' and why it cannot be analyzed or defined? To treat good as a quality that belongs to an object or state of affairs in a simple, unanalyzable fashion would seem to deny that good is relational in any way: Such a position demands that good is not and cannot be *good for* anyone or anything. This would seem to be exactly the sort of position that mistakes intrinsic value for *inherent* value – a view that, when it came up very early in this argument, I considered so incoherent that I relegated my rejection of it to a footnote (see footnote 9 on p.24). Moore's having presented me with a specific example of such a view gives me a better basis for explaining why the view is fundamentally flawed.

Moore's own example of another simple, undefinable quality – the color 'yellow' – reveals the flaw in his understanding by analogy: Moore is of course right that defining 'yellow' in terms of wavelengths of light and cone receptors on the retina and such would not define it adequately – or at all, really. So? The 'yellow' Moore is clearly talking about here, that which cannot be explained or defined for someone who has not experienced it, is neither a quality of light nor a quality of objects that reflect light. It is not in fact a quality of anything at all – it is a *qualia*, which is rather different. The *qualia* 'yellow' (or any other *qualia*) is not a property of any one isolated thing, not even a visual experience: The word *qualia* denotes the subjective phenomenal features of sense experiences, and experiences cannot be defined or explained strictly in terms of what triggers the experience because *experiences are necessarily relational* – there are no experiences without experiencers, just as there are no values without valuers.

So, *contra* Moore, there are in fact parts which compose the whole 'yellow' and the whole 'good.' Moreover, explaining the properties and relations of the parts is a meaningful and useful exercise even if it cannot meet the absurd standard Moore sets. Here, I refer to Moore's claim that one "cannot, by any manner of means, explain to anyone who does not already know it, what yellow is." [Moore, 1929, p.7] No, of course not. But how is one human being defining, analyzing, or explaining 'good' for the benefit of other humans remotely like a sighted person explaining 'yellow' to a blind person? The fact that understanding an explanation or definition of some concept requires some relevant capacity or experience belonging to the one who understands it is not any reason to conclude that the concept is inexplicable or undefinable in some general sense. If I explain 'yellow' with reference to the wavelength of light and the neurophysiology of vision, I have in fact explained the causes of experiences with the 'yellow' phenomenal quality – and it is no less an explanation because someone who has never had such an experience could not fully understand it. Similarly, if I have explained the origin and nature of value in general and moral value in particular – and I think I have – it is no less an explanation because some hypothetical entity which has a mind capable of understanding explanations but has no relevant teleological/value-laden relation to the world (if such an entity is even possible, which I doubt) could not fully understand it. So where exactly is the fallacy, natural or otherwise?

A general statement of the naturalistic fallacy is that an error in reasoning is committed when one draws the conclusion that something is good based on some (natural) properties that belong to it – which Moore takes to be a fallacy due to his opinion that 'good' cannot be defined in any terms, so necessarily it cannot be defined in terms of any properties of those things identified as good. But given that Moore's definition of the naturalistic fallacy hinges on such confused concepts of definition and analysis, I am not sure there is much that is useful in what Moore has to say about the mistakes of other moral philosophers.

For example, Moore accuses John Stuart Mill of equating what is desirable/good with what is desired [Moore, 1929, p.66+], which my analysis of Mill's argument above reveals to be a misreading, and a rather uncharitable misreading at that (see Chapter 2, Section 3 starting on p.33). But even if Mill's argument actually did depend on equating what is desired with what is desirable, that error would properly have been characterized as one of the following: Either Mill is mistaking belief and knowledge, in that desiring something (pleasure) is believing that thing to be desirable, which does not mean that it actually is desirable – which is fallacious, but is not the naturalistic fallacy. Or Mill is drawing an unwarranted value conclusion (pleasure is desirable/good) from a factual premise (pleasure is desired) – which is fallacious, but is not the naturalistic fallacy. Simply defining good as that which is desirable as an end in itself commits no fundamental reasoning error unless one is committed to Moore's insistence that (1) 'good' cannot be defined or explained in any other terms, and/or (2) good is not a relational concept – neither of which seems remotely tenable, for the reasons I outlined above.

While the phrase "naturalistic fallacy" is still in common use, many philosophers and almost all non-philosophers who use it are actually referring to something else: What they intend to refer to is the "is-ought fallacy" I defined in my introduction (see p.7), or more broadly to what I called Hume's problem. I think this confusion arises at least in part because of Moore's criticism of Mill's argument in *Utilitarianism* cited above: One of the reasoning errors that Mill does *seem* to have made is the is-ought fallacy, if one reads his argument as Moore did.

I intended this discussion as background for Moore's Open Question Argument (OQA henceforth), but in fact it is not easy to distinguish the OQA from Moore's definition of the naturalistic fallacy and the confusions on which it is based. In fact, Moore's first articulation of the OQA follows only a few pages after his definition of the naturalistic fallacy, cited above. He presents the OQA as a response to the main alternative to his position – that is, the position that 'good' is not in fact simple and undefinable, but is a complex whole capable of definition and explication:

> **13.** In fact, if it is not the case that 'good' denotes something simple and indefinable, only two alternatives are possible: either it is a complex, a given whole, about the correct analysis of which there could be disagreement; or else it means nothing at all, and there is no such subject as Ethics...
>
> (1) The hypothesis that disagreement about the meaning of good is disagreement with regard to the correct analysis of a given whole, may be most plainly seen to be incorrect by consideration of the fact that, whatever definition may be offered, it may always, be asked, with significance, of the complex so defined, whether it is itself good. [Moore, 1929, p.15]

The first thing one notices about this argument is the problematic prejudice towards intuitionism built into it: The mere fact that there could be disagreement about the correct analysis of a complex conception of the good is taken as reason to reject the idea that 'good' denotes anything complex – which in turn presumes that intuitions about simple, indefinable concepts are never subject to disagreement. The latter is an assumption which Moore does not (and in all likelihood could not) defend.

But let us look at the Open Question Argument itself: For any proposed definition or analysis of the good in terms of some other quality belonging to actions or state of affairs – say, for example, the core utilitarian claim that actions which cause pleasure or alleviate pain are desirable in themselves (i.e. good) – it is always possible for someone who understands moral discourse to grant the definition and still intelligibly ask whether or not that definition is itself good, which calls the definition into question. Suppose I grant that pleasurable actions are good: Thus, when I think that **A** is pleasurable, I am simultaneously thinking that **A** is good. But if for any **A** it is still an open question whether or not it is good that **A** is good, then the definition of good in terms of pleasure is problematic. Thus, for utilitarianism, if **A** is pleasurable, **A** is therefore good. But for every **A**, is it good that **A** is pleasurable? Having an illicit romantic affair would quite plausibly be pleasurable, for example, but it is fairly clear that it is *not* good that having an illicit romantic affair is pleasurable.

However, it is not obvious how much logical force the OQA actually has. The example I presented above seems a bit too easy and obvious, and the examples Moore presents in *Principia Ethica* have that same quality: The characterization of the utilitarian definition of the good given above is, like Moore's characterization of Mill's argument for utilitarianism from which it is drawn, something of a caricature. The utilitarian analysis of the good does not in fact reduce to individual momentary pleasures: The actual normative claim at the foundation of utilitarianism – the claim that happiness (pleasure and the absence of pain, which I won't repeat henceforth), *without respect to whose happiness it is*, constitutes an end in itself – is a much more sophisticated and broad analysis of the good that does not seem so easily dismissed by the OQA.

On reflection, it is not even obvious how one would apply the OQA to this more accurate representation of the utilitarian analysis of the good: Since it is happiness in an abstract, non-proprietary sense which is of intrinsic value, how can one intelligibly ask whether a particular case satisfies it? That is, if one asks "Is it good that (some particular action which advances happiness in the non-proprietary abstract sense, i.e. the general happiness) is good?" it is quite ambiguous whether one is actually challenging utilitarianism's analysis of the good or just utilitarianism's derivation of the right (prescriptive claim) based on its conception of the good (normative claim). In the latter case, it might be the derivation which is problematic rather than the conception of the good. But if one instead applies the OQA in a very general fashion and asks "Is it good that advancing happiness in an abstract, non-proprietary sense is good?" I do not see why one would conclude that the question is still clearly open unless one had *some other basis for disagreement* with the utilitarian perspective on value. I, for example, am not at all convinced by Mill's claim that pleasure (and the absence of pain) is the sole state of affairs desired as an end-in-itself by all humans: However, the OQA does not of itself give me any reason for that doubt about the foundation of utilitarianism – although it could expose the existence of doubts which might otherwise go unnoticed.

This last point, I think, explains something important about the Open Question Argument. Given the criticisms of the OQA I have made here, and many others that have been formulated over the course of more than a century, one might wonder why so many philosophers bring the OQA up repeatedly and seem to take it seriously. One reason to keep the OQA in one's philosophical toolbox is that any claim about the good which cannot survive confrontation with it should certainly trigger closer examination of and suspicion about its justifications. The OQA may not of itself be a knock-down argument against any plausible conception of the good, but it does seem useful as a gatekeeper to rule out the obviously implausible conceptions of the good (such as the trivialized version of utilitarianism's foundation mentioned above), and perhaps also to draw attention to potential problems in more plausible conceptions of the good.

Whether my criticisms of the naturalistic fallacy and the OQA are correct or not, I will attempt to apply the OQA to my own analysis of the good out of respect for whatever motivates philosophers to keep taking it seriously. However, it is not immediately obvious how I might phrase my analysis of 'good' in any neat, formulaic fashion amenable to posing the open question. My most succinct summary of my complex definition of good thus far is almost more poetic than explanatory: Humanity's shared *telos* is the creation/reproduction of culture. Since a *telos* constitutes the highest good for that which has it as a *telos*, humanity's collective highest good is the creation/reproduction of culture.

So, is it good that humanity's collective highest good is the creation/reproduction of our culture? I think the mere fact that someone might consider this question open or unresolved would not be an important criticism of my position: At most, it would be an occasion to look more closely at the specific arguments I made to reach that conclusion – which I welcome.

Section 2: The path ahead

While actually making any strides towards developing an ethical theory on the basis of this foundation of evolved value is beyond the scope of this project, I would be remiss if I made no indication of what potential I see in it after spending so much effort developing it. For the most part, I consider the primary virtue of my approach the rigorous justification it offers for the foundation of an ethical theory, which has been addressed thoroughly above. However, I also think it has some other advantages, either in itself or in comparison with other ethical theories.

Aside from offering firmer normative foundations for prescriptive claims, grounding ethical theory in the increasingly rich and productive science of human nature offers greater explanatory power with respect to several aspects of human moral behaviors and intuitions. For example, my approach provides a framework for addressing the persistent criticism that abstract approaches to ethical theory such as utilitarianism and deontology fail to respect the powerful intuition that we owe things to family and friends that we don't necessarily owe to any human as such. While my foundational normative claim also insists on the value of all humans (as participants in the production and exchange of culture), my account of the more limited moral universes shy of fully non-proprietary universal morality explains the intuition and gives some real weight to family and community obligations. After all, no plausible ethical theory declares that self-interest is wrong, only that it must be circumscribed by consideration for the interests of others – and my perspective treats expanded self-interest within the scope of a moral universe such as a family or social group the same way.

Likewise, I think my evolutionary approach to ethical theory better explains immorality than most ethical theories. No ethical theory denies the existence of drives towards unethical behavior: Mill acknowledges that the (proprietary) desire to pursue happiness for oneself is innate, but that seeing the general happiness as desirable is learned. Kant's rational will often must act in opposition to our inclinations, pitting the categorical imperative against the hypothetical imperatives of self-interested prudence. Aristotelian moral agents must overcome weakness of will and the bad fortune of poor upbringing or desperate circumstances. An evolutionary approach to ethical theory explains both the foundation of ethical behavior and the drive towards unethical/selfish behavior as springing from the same basic source, our reproductive *telos* in its evolved proprietary and non-proprietary aspects. I think this theoretical unity presents a more coherent vision of human nature than any of the rivals discussed above.

Another aspect of ethical theory that might be enriched by my perspective is the conception of moral agency. Thus far in my argument, I have discussed the "what" and "how" of ethical theory, but have for the most part avoided the "who." That is, I have deliberately skirted issues related to both the subjects and objects of ethical theory: By "subjects of ethical theory," I mean moral agents – those beings to whom the *oughts* of ethical prescriptive claims apply. By "objects of ethical theory," I mean the recipients of actions covered by those prescriptive claims. The subjects and objects of ethical theory always overlap, but they are not necessarily identical groups within any ethical theory. For example, utilitarian moral agents are not simply beings who experience pleasure and pain, but rational beings capable of recognizing the intrinsic, non-proprietary value of the happiness of others and thus of recognizing the obligatory nature of the principle of utility: However, the principle of utility prescribes actions towards any being that can experience pleasure and pain, meaning that animals are objects for utilitarian ethics even if only humans (as far as we know) are subjects.

One reason for avoiding these issues is simply that they were not directly relevant to my argument. Whatever theoretical allowances philosophers may include to the contrary – for example, Kant's discussion of God or angels as beings with a rational will but no inclinations to be overcome by reason [Kant p.215 (4: 414)] – there is little dispute at the level of ethical theory that the primary subjects/moral agents of concern are humans. Similarly, the primary objects of concern are humans, with any additional objects determined by the specific nature of the highest good/primary normative claim which serves as the foundation for that particular ethical theory: The utilitarian highest good of happiness, defined simply in terms of pleasure and avoidance of pain, implies the need to include other sentient beings as objects of primary moral concern – while the Kantian highest good of the rational will does not seem to include animals as objects of moral concern in any immediately obvious way.

Since my motivation for examining the details of particular ethical theories was simply to confirm that they instantiated the general solution to Hume's problem I proposed, what mattered was that they were in fact grounded in a claim about what is valuable to humans, not whether that same thing might also happen to be valuable to something other than humans. However, being rooted in biology, my evolutionary approach to ethical theory would seem to address the issue of non-human moral subjects and objects very naturally. Since the basis for the non-proprietary value foundations of evolutionary ethics is participation in the production and exchange of culture, any creature which can potentially do so automatically becomes a part of universal morality rather than residing in a separate moral universe. While it may seem like a somewhat abstruse science-fictional sort of concern, I think it is interesting and potentially important that the basis for morality I propose does not just encompass *homo sapiens*, but any living creature (i.e. any being organized around the reproductive *telos*) which is capable of participating in the production and exchange of culture. Perhaps SETI will someday open communication with aliens; or perhaps we will discover that dolphins are even cleverer than we think and we will learn how to interpret their language (or they ours)? In the event, I have an ethical theory that can accommodate the ethics of interspecies relations.

The primary obstacle I see for an ethical theory rooted in the evolutionary foundation I have developed here is simply that I currently don't even know where I would start. The teleological foundation of my approach gives the development of an evolutionary version of Aristotelian virtue theory a natural appeal, and the core value of participation in the production and exchange of ideas would seem to imply a strong parallel with Aristotle's virtues of intellect as well as his virtues of character. However, there is also a respect in which some consequentialist prescription to advance the common good is implied by my emphasis on culture: But how that might play out as an ethical principle is not as obvious as the principle of utility – which is not necessarily a bad thing, given that the principle of utility's limitations and flaws are widely acknowledged. In any event, given the many possibilities available, perhaps any future ethical and/or political theories rooted in my proposed foundation should be left where they currently reside – the future.

REFERENCES

Aristotle. *Nicomachean Ethics*, translated by Joe Sachs. Newburyport, MA: Focus Publishing, 2002. Parentheses within reference brackets for all Aristotle citations indicate the page, column, and approximate line numbers of the standard nineteenth century Bekker edition of Aristotle's works in Greek on which the translation is based. Citations lacking parenthetical numbers indicate a reference to commentary rather than translation.

Aristotle. *On the Soul*, translated by J. A. Smith, in *The Complete Works of Aristotle, Volume One*, edited by John Barnes. Princeton: Princeton University Press, 1985.

Bickerton, Derek. *Language and Human Behavior*. Seattle: University of Washington Press: Seattle, 1995.

Brosnan, S. F. and de Waal, Frans B.M. "Monkeys reject unequal pay" *Nature*, Vol. 425, No. 6955, 297–299. September 18, 2003.

Charnov, Eric L. and Krebs, John R. "The Evolution of Alarm Calls: Altruism or Manipulation?" *The American Naturalist*, Vol. 109, No. 965, 107-112. January-February, 1975.

Coyne, Jerry A. *Why Evolution Is True*. New York: Viking Penguin, 2009.

Darwin, Charles. *On the Origin of Species (A Facsimile of the First Edition)*. Cambridge: Harvard University Press, 2001.

de Waal, Frans B.M. "The chimpanzee's sense of social regularity and its relation to the human sense of justice" *American Behavioral Scientist*, Vol. 34, No. 3, 335–349. January, 1991.

Gould, Stephen Jay. *The Structure of Evolutionary Theory*. Cambridge: Harvard University Press, 2002.

Hamilton, W.D. "The Genetical Evolution of Social Behaviour" *Journal of Theoretical Biology*, Vol. 7 No. 1, 1–52. July, 1964.

Hume, David. *A Treatise of Human Nature*, edited by David Fate Norton and Mary J. Norton. New York: Oxford University Press, 2000.

Kant, Immanuel. *Groundwork for the Metaphysics of Morals*, translated by Arnulf Zweig, edited by Thomas E. Hill, Jr. and Arnulf Zweig. Oxford: Oxford University Press, 2002. Parenthetical numbers used in all references to this work indicate the volume and page number of the standard Königlich Preußische Akademie der Wissenschaften edition of Kant's collected writings (*Kants gesammelte Schriften*) on which the translation is based.

Lennox, James. *Aristotle's Philosophy of Biology*. Cambridge: Cambridge University Press, 2001.

Lewin, Roger. "Accidental Career" *New Scientist*, Vol. 63, No. 909, 322-325. August 8, 1974.

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

Mayr, Ernst. *The Growth of Biological Thought*. Cambridge: The Belknap Press of Harvard University Press, 1982.

Mill, John Stuart. *The Basic Writings of John Stuart Mill: On Liberty, The Subjection of Women & Utilitarianism*, introduction by J. B. Schneewind, notes and commentary by Dale E. Miller. New York: Modern Library, 2002.

Moore, George Edward. "Is Goodness a Quality?" reprinted in *Philosophical Papers*, pp. 89-101. London: Allen & Unwin, 1959.

Moore, George Edward. *Principia Ethica*. Cambridge: Cambridge University Press, 1929.

Plato. *Republic*, translated by Robin Waterfield. Oxford: Oxford University Press, 1993.

Range, Friederike; Horn, Lisa; Viranyi, Zsófia; and Huber, Ludwig. "The absence of reward induces inequity aversion in dogs" *Proceedings of the National Academies of Science*, Vol. 106, No. 1, 340-345. January 6, 2009.

Russell, Bertrand. *The Problems of Philosophy*. London: Williams & Norgate, 1912.

Sachs, Joe. *Aristotle's Physics: A Guided Study*. New Brunswick: Rutgers University Press, 1995.

Sayre-McCord, Geoffrey. "Mill's 'Proof' of the Principle of Utility: A More than Half-Hearted Defense" *Social Philosophy & Policy*, Vol. 18, No. 2, 330-360. Spring 2001.

Searle, John R. "How to Derive 'Ought' From 'Is'" *The Philosophical Review*, Vol. 73, No. 1, 43-58. January, 1964.

Sober, Elliot and Wilson, David Sloan. *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge: Harvard University Press, 1998.

Trivers, Robert L. "The Evolution of Reciprocal Altruism" *The Quarterly Review of Biology*, Vol. 46, No. 1, 35-57. March, 1971.

Tylor, Edward B. *Primitive Culture: Researches into the Development of Mythology, Philosophy, Religion, Art, and Custom*. Boston: Estes & Lauriat, 1874.

Whiten, Andrew; Spiteri, Antoine; Horner, Victoria; Bonnie, Kristin E.; Lambeth, Susan P.; Schapiro, Steven J.; and de Waal, Frans B.M. "Transmission of Multiple Traditions within and between Chimpanzee Groups" *Current Biology*, Vol. 17, p.1038–1043. June 19, 2007.

Wilkinson, Gerald S. "Food Sharing in Vampire Bats" *Scientific American*, Vol. 262, No. 2, p.76-82. February, 1990.

Wilson, David Sloan. "A Theory of Group Selection" *Proceedings of the National Academy of Sciences*, Vol. 72, No. 1. January, 1975.

Wilson, David Sloan. *Darwin's Cathedral: Evolution, Religion, and the Nature of Society*. Chicago: The University of Chicago Press, 2002.

Witkowski, Ken. "The 'Is-Ought' Gap: Deduction or Justification?" *Philosophy & Phenomenological Research*, Vol. 36, No. 2, 233-245. December, 1975.