TOWARDS A CONSTRUCTIVIST EUDAEMONISM

Robert H. Bass, Jr.

A Dissertation

Submitted to the Graduate College of Bowling Green
State University in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2004

Committee:

Edward McClennen, Advisor

Catherine Cassara
Graduate Faculty Representative

Fred D. Miller, Jr.

Loren Lomasky

ABSTRACT

Edward F. McClennen, Advisor

Eudaemonism, I take to be the common structure of the family of theories in which the central moral conception is *eudaemonia*, understood as "living well" or "having a good life." In its best form, the virtues are understood as constitutive and therefore essential means to achieving or having such a life. What I seek to do is to lay the groundwork for an approach to eudaemonism grounded in practical reason, and especially in instrumental reasoning, rather than in natural teleology. In the first chapter, I argue that an approach based in natural teleology will not work. In the second, the claims of decision theory to be an adequate formal representation of instrumental reasoning are examined and found wanting. In the third, I develop an account of ordinary instrumental reasoning. In the fourth, I discuss the structure of eudaemonism, with the aim of showing that there is an intelligible and attractive doctrine that can be disentangled from the natural teleology. In the fifth, I sketch an argument showing that instrumental reasoning, as explicated in the third chapter, can bear on the selection of final and ultimate ends, and that it is plausible that the instrumental approach to moral theory that I am urging yields conclusions with a eudaemonistic structure. I also indicate directions for further development and exploration.

ACKNOWLEDGMENTS

For any dissertation, though one person may be listed as the author and receive the degree, many more are involved in providing essential support, financial, intellectual and personal. This dissertation is no exception. I'd like to take the opportunity to express my deep gratitude to the graduate faculty of the Philosophy Department of Bowling Green State University and especially to the members of my committee, E. F. McClennen, Fred Miller and Loren Lomasky. Intellectually and personally, they have all profoundly influenced me, and the dissertation is better for their contribution. In addition, I'd like to acknowledge the help and support I received from Mike Robins, who would have added his contribution to the dissertation committee, and no doubt in other ways, had he not tragically died in 2001. He was a good man, a good friend, and an indefatigable intellectual sparring partner. He will be missed. My deepest debts, however, are personal, to family, friends and loved ones who have provided ongoing support of many kinds over what sometimes seemed to be an endless process. Deserving of special mention are my parents and, more than anyone else, my wife, Gayle Dean.

TABLE OF CONTENTS

INTRODUCTION: CONSTRUCTING EUDAEMONISM

That eudaemonism is an attractive structure for a moral theory is attested both by its adoption by many of the ancients[1] and by much of the contemporary interest in virtue ethics. There is a problem, however, in that the ancient theorists often tied their eudaemonism to a form of natural teleology which is certainly not acceptable in detail now, while modern work in the field is often subject, if not to guilt, then at least to suspicion, by association. Moreover, confirmation for the suspicion may readily be found in the work of contemporary eudaemonists who adhere to or seek to rehabilitate (perhaps in an improved form) the ancients' natural teleology.[2]

In the current project, I attempt to move beyond this and toward a version of eudaemonism independent of its ancient moorings in natural teleology. More specifically, I seek to move toward a *constructivist eudaemonism*, which will bring

[1] Among the ancient eudaemonists are Aristotle, Epicurus and the Stoics. In fact, among the ancient Greeks, only the Cyrenaics were *not* eudaemonists of some stripe. (Annas 1993)

[2] E.g., Arnhart 1998, Irwin 1980, Rasmussen and Den Uyl 1991.

together and rely upon theses within three areas of long-standing philosophical interest to me, *eudaemonism*, *constructivism* and *instrumental reason*. Together, these constitute the background against which what I am attempting should be understood. In the way of brief explanation, I offer the following.

First, by *eudaemonism*, I refer to the common structure of the family of theories in which the central moral conception is *eudaemonia*, understood as "living well" or "having a good life."[3] In the form I take to be best, and which I shall therefore highlight, the virtues are understood as constitutive means to achieving or having such a life.[4] Though I prefer "eudaemonism" as a label, the position has a close affinity, sometimes amounting to identity, with what is commonly called "virtue ethics" or "perfectionism."

Though the structure of eudaemonism is appealing, it needs to be separated and considered apart from the traditional grounding of eudaemonism in natural teleology – that is, in ends, purposes or goals that are supposed to somehow be given to us "by nature." In my view, for ethics, natural ends are a dead end. (Nor am I satisfied with Hurka's 'intuitive appeal is enough.'[5]) In part, the reason for disconnecting eudaemonism from theories of natural ends is to avoid the guilt or suspicion by association mentioned above. But also, it is important to see that there is an intelligible and attractive doctrine that *can* be separated from the natural teleology, that eudaemonism does not stand or fall with the fortunes of natural teleology.

Second, I find *constructivism* a plausible account of what we mean or should

[3] I think it misleading, without further explanation, to employ the traditional translation of "eudaemonia" as "happiness."

[4] This is still not sufficient to distinguish eudaemonism, or my favored version of it, from all other moral theories, but I will not attempt anything more complete here. A fuller characterization will occupy a substantial portion of Chapter Four, "The Structure of Eudaemonism."

[5] Hurka 1993, 28-33.

mean by ethical objectivity. The question to which constructivism provides one (kind of) answer is, "What is it for an ethical claim or judgment to be true or correct or justified?" Constructivism offers what might be called a *practical-reason-first* account of moral objectivity. This is best approached by contrasting it with two other possible answers. On one hand, there are substantive moral realists[6] who think that the correctness of moral claims depends upon the existence of moral facts somehow "out there." The moral facts pertain to the existence or instantiation of moral or value-properties, to what is right, good or valuable and to the relations between these facts. On a substantive realist's view, a correct moral claim is one that gets things right about the moral facts. In general, the substantive realist thinks both that there are such facts and that we have some kind of cognitive access to them.[7] At the other extreme are those who may be termed moral skeptics or nihilists[8] who, in one way or another, deny that any moral claims are true or correct or, at least, that there can be any knowledge of their truth or correctness. Interestingly, the skeptics typically share the same model of moral truth or cognition as

[6] I borrow the term from Christine Korsgaard:

> There is a trivial sense in which everyone who thinks ethics isn't hopeless is a realist. I will call this *procedural* moral realism, and I will contrast it to what I will call *substantive* moral realism. Procedural moral realism is the view that there are answers to moral questions, that is, that there are right and wrong ways to answer them. Substantive moral realism is the view that there are answers to moral questions *because* there are moral facts or truths, which those questions ask *about.* (Korsgaard 1996b, 35. See also surrounding discussion, 34-37.)

Constructivists differ from substantive moral realists not in whether they accept that there are correct moral claims but in how they understand the correctness of moral claims.

[7] Technically, these are distinct assumptions. In principle, one could be a moral realist in the sense of believing that there are moral truths, without being a cognitivist – that is, without believing that we have any way of knowing what the moral truths are. Tristram Engelhardt, if I understand him correctly, holds this view. But the realist claim, that there are moral truths, and the cognitivist claim, that we have some way of knowing what is morally right, are so regularly accepted or denied together, that it is for practical purposes sufficient, on one hand, to call someone a realist or a cognitivist or, on the other, to call her a non-realist or non-cognitivist, to indicate her position on both questions.

[8] Harman 1977; Harman, in Harman and Thomson 1996; Mackie 1977. Others with different terminological preferences may call such theorists subjectivists, relativists or non-cognitivists. Since the terminology is unsettled, not all who are described or self-described by one of these terms would fit within the parameters of my definition.

the substantive realists. Both think of moral claims as being correct or incorrect in virtue of their relation to independent moral facts.[9] The moral facts are truth-makers for correct moral judgments. The constructivist's approach is different. He agrees with the substantive moral realists that the skeptics are mistaken (there *are* correct or justified moral claims) and with the skeptics that the mysterious properties or entities to which the substantive realists appeal don't exist.[10] Instead, he holds that we can identify correct moral reasoning – or better, correct practical reasoning – at least to the extent of being able to recognize better and worse instances of such reasoning. In substantive realist theories, correct moral reasoning is reasoning that tracks or tends to track the moral facts; in constructivist theories, the order of dependence is reversed: what is morally correct is whatever is picked out by correct moral reasoning.[11]

Above, I indicated that it was better to think of the constructivist as focusing upon practical reasoning rather than just upon moral reasoning. This is because I take practical reasoning to be a broader classification than moral reasoning, and if correct practical reasoning is the constructivist's focus, then correct moral reasoning should be understood as a special case. Briefly, practical reasoning is, in the first instance, reasoning about what to do, with a range from the trivial to the momentous, from whether to scratch an itch to whether to fight in a war. Moral issues tend to be clustered toward the momentous

[9] More generally, I think the same line of thought is often behind relativism or subjectivism. The model remains the same: there is a comparison or matching between, on the one hand, behaviors or beliefs, and on the other, standards, but the only standards that seem (to the relativist or subjectivist) to be available for comparison are social or personal.

[10] Or, if they do exist, they are epistemically inaccessible to us and therefore useless for the guidance of our deliberation or action.

[11] Though I think that much can be said on behalf of the coherence and plausibility of constructivism as a way of understanding moral objectivity, I don't intend to devote much space to directly and abstractly defending it. That is why there are chapters addressing both eudaemonism and instrumental reasoning but not one on constructivism. The best argument that we can make progress in moral theory along constructivist lines consists of progress made. That is what I hope to provide.

end of the spectrum and often share further features.[12] If I am correct in thinking of moral reasoning as a special case of practical reasoning, there is no room for the suggestion that something might be practically correct but morally wrong or morally correct but practically wrong. I take the two, moral and non-moral practical reasoning, to be continuous and the distinction between them to be fuzzy (but not, for that reason, a non-distinction).[13]

Returning to the main line of discussion, if the constructivist's project is to be carried through, some account of correct practical reasoning will clearly be needed. Ideally, this account should itself be either uncontroversial or readily defensible, which brings us to the third area of philosophical interest mentioned above, for I wish to suggest that a promising place to begin is with the obvious power and normative force of *instrumental reasoning*. Many recent thinkers have been similarly inclined,[14] but, at least as often, proposals to begin developing a moral theory grounded in instrumental reasoning have been met with skepticism, generally centering around claims that a correct moral theory must have some bearing on the ends that we ought to pursue, not just address questions about the most effective or efficient ways to pursue given ends.[15] I

[12] Among these are the non-overridability of moral requirements by other, non-moral, considerations, the fact that we take ourselves and expect others to have reasons for holding moral positions, and that there is some kind of requirement of impartiality. Though these are characteristic, I think that, neither singly nor in combination, do they provide a set of necessary and sufficient conditions for something to count as a moral issue. All seem subject to counter-examples such that either some issue that would normally be classified as moral fails to meet the conditions or some issue not normally so classified succeeds (or both).

[13] Note that if there is a distinction between moral and non-moral practical reasoning, it is at least in principle possible that correct practical reasoning can be identified, as the constructivist maintains, but that no correct practical reasoning leads to recognizably moral conclusions and therefore that, contrary to what I claimed above, moral reasoning would not be a special case of practical reasoning. In that case, I think we would do better to say that practical reasoning does not lead to anything that can properly be called a *moral* theory. Perhaps our practical reasoning will turn out to be too deeply infected by contingent differences in the starting points from which agents begin their reasoning.

[14] E.g., Schmidtz 1995, Gauthier 1986.

[15] E.g., Piper 1986.

agree with the critics that moral theory must bear on the correctness of ends, but find unconvincing the claim or argument that instrumental reasoning cannot do so. At one time, I thought (or hoped) that *everything* needed for moral theory could be done in terms of instrumental reasoning. I no longer think so: if we take instrumental reasoning seriously enough, we will be forced beyond it.[16] However, it remains interesting and, I think, fruitful to see how far we *can* go, starting from instrumental reasoning.[17]

The general shape of the view that I am trying to work towards is that, due to certain pervasive features of human life and action, especially features having to do with conflicts between or among goals, people have reasons to acquire systems of goals that have the kind of structure recommended by eudaemonism, that is, in which there is an over-arching goal of living well or having a good life and to which the virtues are constitutive means. The argument I shall present unfolds in several stages, and it may not always be obvious how the different pieces are meant to fit together. To make matters easier, what follows is a sketch of the main phases of the argument.

The first two chapters are aimed at ground-clearing. Each takes as its target a prominent theory and seeks to exhibit its inadequacy for the purposes of the current project. By implication, there is room for and need for alternatives.

In the first chapter, "The Insufficiency of Natural Ends," I focus upon natural

[16] In other words, in taking instrumental reasoning as a starting point, I do not mean it to be also a stopping point. However, if, as I believe, there are ways in which practical reasoning extends beyond the instrumental, I will not rely upon them.

[17] If instrumental reasoning pushes us beyond itself, does that provide us with a sense in which we can, after all, do everything needed for moral theory in terms of instrumental reasoning? Not necessarily. Dialectical pressures internal to our understanding of instrumental reasoning may lead us to recognize a place for non-instrumental practical reasoning without fully specifying the form or content of that non-instrumental reasoning.

teleology[18] – the idea that there are ends or norms somehow set for us by nature – with the objective of dissociating eudaemonism from these traditional moorings. The attempt is guided by two thoughts: First, on an acceptable account of what natural teleology consists in, it will turn out not to be satisfactory for the purposes of moral theory. We *can* make sense of natural teleology, but on the best understanding, it is unhelpful for ethics: On the face of it, it entails counter-intuitive consequences, and, even if those are avoided, appears to deliver inapplicable prescriptions. Moreover, however it is understood, it confuses explanation with justification and fails to address genuine moral perplexity.

For these reasons, eudaemonism should not, if we can avoid it, be identified or inseparably associated with natural teleology, for so associated it can be shown to be untenable. Naturally, showing eudaemonism to be untenable when tied to natural teleology is not equivalent to showing it to be tenable once that link is broken. Nonetheless, and this is the second point, it is an important preliminary. With natural teleology dismissed or set aside, we will be better placed to see what eudaemonism *itself* is and involves and to assess eudaemonism without being distracted by the various debates that appeals to natural teleology draw in their train.

In the second chapter, "Decision Theory and Instrumental Reasoning," I consider the nearly canonical treatment of instrumental reasoning provided by standard decision theory. My principal concern has to do with its use in a normative role, to explicate what agents have reason to do, given their preferences. I argue that decision theory is

[18] Not only eudaemonists but some contemporary explorers of what has come to be called "evolutionary ethics," are also inclined to think that some grounding of their theories in natural ends or functions is workable.

unsatisfactory in this role, in that it relies upon assumptions about preferences which finite agents are not in a position to satisfy, and ultimately, if taken as a general account of instrumental rationality, leaves us unable to make sense of the normative distinction between ends and means. It is better cast in a supporting role.

The remaining chapters are more constructive. In the third, "The Scope of Instrumental Reasoning," I present an account of ordinary instrumental reasoning, which is, in the first place, more modest in its ambitions than standard decision theory, in that it does not aspire to be fully formalizable. In the second, it makes more modest demands upon agents, only presupposing capacities for thought, deliberation and comparison of options that appear to be within our grasp. However, the account is not merely decision theory minus something: it has independent interest. In particular, beginning from the simplest case of instrumental reasoning, the selection of some means for the sake of its causal contribution to bringing about some single objective, the account readily generalizes to cover cases in which multiple objectives or goals have a bearing upon action, to cover cases in which decision-making among some set of options can only be rationalized by appeal to some further combinatorial principle, and to cover constitutive reasoning in which the means adopted at least partially constitutes the objective for the sake of which the means is adopted. Finally, though it does not depend especially upon the account of instrumental reasoning offered here, I offer a partial explanation of the normative force of instrumental rationality that shows that it need not depend entirely upon some normative force or value attached to the ends from which the reasoning proceeds.

The fourth chapter, "The Structure of Eudaemonism," has two principal aims. One is to secure the plausibility of the claim that eudaemonism is an appealing structure for a moral theory. Unfortunately, I have no idea how to prove appealingness.[19] In the end, the answer to the question whether a theory is appealing must be left to those considering it, and whether it is appealing depends on whether they find it so. Still, something may be done to clear away obstacles and to disentangle the theory from accretions or misunderstandings that may stand in the way of a fair assessment of its appeal. If we want to consider whether a theory is appealing, we need to make sure that we are considering *that* theory rather than something else or some amalgam of that theory with something else.

Accordingly, my second concern is to more directly characterize the principal features of eudaemonist theories. Now, what is common to all eudaemonist theories, the normative centrality of living well or having a good life, can, with only a bit of ingenuity, be construed to apply to virtually any other moral theory as well.[20] But adding conditions to rule out non-eudaemonist theories does not help, for the plausible candidates for additional conditions would also exclude theories, such as Epicureanism, that clearly belong within the eudaemonist camp. Naturally, this makes it difficult or impossible to produce a characterization of eudaemonism in the form of some illuminating set of necessary and sufficient conditions. What I shall try to do instead, after brief attention to some further possible misunderstandings, is to develop further an account of the structural features of eudaemonism at its best, where eudaemonia is understood as an inclusive ultimate end of living well to which the moral virtues are constitutive means.

[19] I do not of course mean that intuitive appeal is decisive for the correctness of a moral theory or that the correct theory cannot be counter-intuitive in various ways. But that does not mean it has no evidential value whatever. If, to pursue a legal metaphor, intuitive appeal does not settle the case in favor of a moral theory, it still may be relevant, first, to getting the theory a hearing in the first place, and second, to establishing a (rebuttable) presumption in its favor.

[20] Other theorists may not, however, think that the most natural or illuminating way to describe their positions.

This will also require further elaboration of the ways that ends may be classified and related to one another as well as some account of the way that the virtues fit into the eudaemonist framework.

In the final chapter, "Reasoning About Ends," I try to bring together what has been developed in earlier chapters to show the bearing of instrumental reason upon the selection of ends, including ultimate ends. Some decision theorists and partisans of instrumental reasoning can be expected to object that instrumental reasoning takes ends as given and can only address questions about the relative efficacy of means to given ends, but can have no bearing on the correctness of the ends themselves.[21] I believe this is a mistake. Abstractly, the relevant point can be put like this: Instrumental reasoning can bear on the correctness of ends if the selection or adoption of some end can itself be a better or worse means to some other end or ends. More concretely, this can be illustrated by the kind of motivational change that an agent may undergo in breaking a bad habit. The agent may conclude, on the basis of his existing corpus of ends, that giving up the habit (where that involves actually changing the set of ends that he seeks) will better serve his existing corpus of ends and that whatever costs are attendant upon making the change are less than the benefits to be expected. Once he has successfully made the change, he will have a somewhat different corpus of ends.

I think examples like that are sufficient to show that instrumental reasoning *can* bear upon the selection of ends, but it is plain that much more needs to be done, especially if I am to defend the claim that it bears upon the selection of final and ultimate

[21] For example, Bertrand Russell writes, "'Reason' has a perfectly clear and precise meaning. It signifies the choice of the right means to an end that you wish to achieve. It has nothing whatever to do with the choice of ends." (1955, vi)

ends. How instrumental reasoning could bear upon selecting ends which are not themselves means to anything further might still reasonably be thought to be problematic, even when it is granted that such reasoning can bear upon the selection of *some* ends. I argue, however, that that rather tricky passage can be negotiated and further, that we end up not only with some structure or other including ultimate ends and others related as means or constituents, but that the structure is plausibly that of eudaemonism – i.e., that there will be an ultimate end, which can be characterized as one of living well, and that the virtues will figure as constitutive means.

Though I intend to limit myself to making a case for the plausibility of this thesis about the bearing of instrumental reason upon ultimate ends, a number of issues invite further exploration, and, at the end of the chapter, I briefly discuss some. Of special interest to me are questions about the scope of the audience addressed by the kinds of considerations I offer and what that implies or suggests about moral education, about non-instrumental practical reason and the possibilities for a eudaemonist or virtue-ethical approach to politics.

CHAPTER ONE: THE INSUFFICIENCY OF NATURAL ENDS

1.0 Introduction

A substantial and enduring tradition in ethics, that of natural law, has conceived itself as finding or discovering norms present in nature.[1] Somehow, there is natural teleology or there are natural ends which determine what our good is. By understanding the natures of things – especially by understanding human nature[2, 3] – we can understand

[1] Wild 1953, Strauss 1953, Korsgaard 1996b. I take it to be essential to the natural law tradition that there is some kind of appeal to nature. I will say something further below about how the appeal is thought to work. As I construe it, it is not sufficient simply to hold that some moral claims are objectively correct or that their correctness does not depend on contract, custom or convention as, for example, Hart seems to do (1984, 77f.). Strauss also at times seems to understand natural law as equivalent to moral objectivity (p. 3).

[2] I do not think deep conceptual or theoretical problems stand in the way of identifying human nature, so I will base no criticisms upon that. It is probably true that we cannot provide an illuminating set of necessary and sufficient conditions for being human, but I think the insistence that we must have such conditions (as distinct from conditions which generally hold for human beings or in human societies) is itself a hold-over from an essentialism about biological species which, in the aftermath of Darwin, is untenable.

[3] Of course, understanding human nature has to include an understanding of human sociality. It

the purposes or goals or principles to which we are suited or which suit us by nature and find guidance as to what we ought to do, what kinds of lives to live, what kinds of characters to cultivate. Importantly, for my purposes, many or most eudaemonists have placed themselves within that tradition, extending at least from the time of Aristotle,[4] through the Stoics and the Thomists to Aristoteleans and neo-Aristoteleans[5] of the present day.[6]

Though some version of the natural law position has probably been held by the majority of eudaemonists, I wish to distance myself from it. I believe it subject to decisive objections and that the appeal of eudaemonism is radically undermined if it is insisted that it is inseparably linked to a philosophically suspect view[7] about the place of

should not be presumed that adequate understanding can be achieved by the examination or investigation of isolated individuals.

[4] According to Julia Annas, "[f]or Aristotle, it is just as naive as it is for us to ask what the point is of a human life. This is not a well-defined question; for there is no well-defined larger system that a human being is part of. So Aristotle does not have a 'universal teleology'; and the teleology that he does have is not a theory about human lives." (1993, 139) Elsewhere, commenting on Aristotle's function argument in Book I of the *Nicomachean Ethics*, she downplays the reference to nature, saying that "it turns up once in verbal form, but somewhat casually." (p. 144)

I find this attempt to minimize the teleological dimensions of Aristotle's ethical thinking puzzling. In the first place, she seems to impose an extravagant condition upon interpreting his thought about the human good as teleological – that we could only do so if we saw human lives as fitting into a larger system. I don't see that Aristotle ever commits himself to this. As John Cooper says, "The good of each species is judged merely on the basis of its single nature, and without assuming it was made for any further purpose." (1996, 280) Second, Aristotle clearly thinks that human beings exist by nature and have natures and that "... action for an end is present in things which come to be and are by nature." (*Physics* 198b1 7-8, cf. 199b1 32) Third, in the function argument in the *Nicomachean Ethics* (hereafter cited as *NE*), he seems to assert just what Annas denies that he does: "[T]o say that happiness is the chief good seems a platitude, and a clearer account of what it is is still desired. This might perhaps be given, if we could first ascertain the function of man.... Or as eye, hand, foot, and in general, each of the parts evidently has a function, may one lay it down that man similarly has a function apart from all these?" (1097b1 22-32) Surely, however the function argument is best understood, Aristotle is not treating it as naive to ask for the point of a human life. (All quotes from Aristotle, unless otherwise noted, are from the New Oxford Translation, 1984.)

[5] Of course, not all who would call themselves Aristoteleans or neo-Aristoteleans identify with the natural law tradition. A notable exception is Alasdair MacIntyre (1984).

[6] Annas 1993; Finnis 1980; Irwin 1980; Machan 1975; Miller 1995; Rasmussen and Den Uyl 1991; Wallace 1978.

[7] My concern is not that the appeal to natural teleology is *scientifically* suspect but rather that it

values in nature.[8]

1.1 The Appeal to Nature

We can distinguish at least three ways in which the appeal to nature was supposed to work: nature as *limit*, nature as *potential* and nature as *direction*. I shall only briefly speak of the first two – in fact, shall treat them together – for what is distinctive about the natural law tradition has been what it has had to say about the third, about the way in which attending to our nature can provide guidance as to what to do.

1.11 Nature as Limit and as Potential

Our nature is, on one level, simply what is inevitable or unavoidable about ourselves. We are limited physically, psychologically, cognitively and motivationally. We have certain traits and not others, and the fact that certain patterns of action and response are not open to us sets limits on what we ought to do.[9] If all of the limits were merely idiosyncratic, though they would be relevant to what particular persons ought to

fails to provide the right sort of answers in ethical questions.

[8] Thomas Hurka's perfectionism is, though clearly related, not a version of eudaemonism as I construe it. Like me, he seeks to dissociate his position from any grounding in natural teleology. I think he is right to do so, but that he has no adequate replacement. The appeal to natural teleology may fail, but it is at least an attempt to answer a real question which might be phrased, "which way is up?" By invoking natural ends, the natural teleologists try to say which directions of change within the scope of possible different exercises of our natural capacities count as improvements and which do not. Hurka, of course, has an account of what counts as improvement, but seems to rest his conclusions on their intuitive appeal. (1993, 28-33) He speaks frequently of realizing or fulfilling or perfecting human nature, but doesn't explain why his candidates count as improvements.

[9] Some patterns of action and response might be motivationally inaccessible to us not because of sheer impossibility but because we are incapable of seeing them as choice-worthy. However, the point needs careful qualification because there may be a developmental story about how we come to see those patterns as choice-worthy. The motivations of the mature person of settled moral character (the practically wise person) may be opaque to the beginner, without its being the case that there is no developmental path that leads from what the beginner sees as choice-worthy to what the practically wise see as choice-worthy.

do or could reasonably aspire to do, there would be little place for attention to them (casuistry apart) in ethics. If, however, some or many of the limits are extremely widespread – if, that is, they can reasonably be said to be matters of human nature rather than matters of the particular characteristics of particular agents – then they may well have an important bearing upon moral theory. Though this might conceivably be contested (in the name of Original Sin, perhaps), it should be relatively uncontroversial for anyone endorsing some reasonably wide-scope version of the Kantian dictum that " 'ought' implies 'can' "[10] and its corollary, " 'cannot' implies 'not-ought.' "[11]

The same point can be deployed in a more positive form. Through attention to human nature, we may not only discern limits but disclose possibilities. We can look at the kinds of lives lived in different communities, different historical periods and in different cultural contexts. We can look for generic features found in almost all lives, however different they may otherwise be, such as engagement in some productive occupation, involvement with family and community and so on. We can also seek and may find that there are ways of life that are widely exemplified and seem to fit into recognizable social roles, such as the lives of a soldier or an artist or an intellectual, rather than involving just the possibly idiosyncratic activities or predilections of individuals.[12]

[10] "[W]hen the moral law commands that we *ought* now to be better men, it follows inevitably that we must be *able* to be better men." (Kant 1960, 46) Also, a character in an example "judges, therefore, that he can do something because he is aware that he ought to do it." (Kant 1997, 30)

[11] I believe we can identify some exceptions to " 'ought' implies 'can'," cases in which it is proper to say that something ought to be or have been done but in which it cannot be or could not have been done. The plausible cases involve culpable inability and various sorts of conflict between obligations. However, these are rare exceptions and, for the vast majority of cases, it is true that if one cannot do it (whatever "it" is), it is not the case that one ought to do it. A discussion of some such cases may be found in Kavka 1986, 309-314.

[12] It is not being assumed that unique or idiosyncratic lives that are not (recognized by us to be) integrated into the societies in which they are lived cannot be expressive in interesting ways of the possibilities of human nature, but rather that widespread ways of life that are integrated into the societies in

Among these, we can ask which ways of life, on the one hand, are found to be satisfying or worthwhile by those who live them,[13] and, on the other, are admired or respected by others. We can also ask whether, among these ways of life judged to be worthwhile, there are any traits of character that are generally common to, distinctive of and regarded as important within those ways of life. This kind of investigation, which can obviously be pursued further than has been sketched here, can suggest a great deal about what kinds of lives may be good – about the range of possibilities for good lives – including drawing our attention to possibilities that would not have occurred to us apart from the investigation. Again, however, recognition of this fact is something that should be acceptable to moral theorists of many sorts. Through considerations of this kind, we may reach a heightened awareness of what may possibly count as a morally good life, but must look elsewhere to select among them or even to be sure that a morally good life *is* among the options we have examined. Though the approach can be described as an appeal to nature, it is not distinctive of a natural law position.

1.12 Nature as Direction

What has been distinctive about the natural law position has been the claim that nature not only (negatively) imposes limits on what can count as a good life or (less negatively) makes different kinds of lives possible, some of which may count as morally

which they are lived have passed a test that the unique and idiosyncratic have not (yet). There is a presumption in their favor as being expressive of interesting possibilities of human nature.

[13] It is at least a plausible initial assumption that morally good lives will typically be found to be satisfying or worthwhile to morally good persons. Enkratic lives, in which there is knowledge of and concomitant action upon what is morally good in the face of inner conflict, cannot be offered as a possible counter-example unless one already has some kind of account of what a good life is. In particular, we would need to know whether the enkratic ultimately count as morally good. But, at this stage of the investigation, that is yet to be provided.

good, but that nature provides direction – that somehow to be found in nature are norms, standards or ends which provide direction towards living good lives. If we properly attend to nature, we find not just limits to what we can do or possibilities for what we may do but what we *should* do. We find goals or ends by which to direct our actions.[14]

How was this supposed to work? The root idea probably came from consideration of artifacts.[15] A good knife is one that is sharp, rust-free, well-balanced, that keeps an edge and so on. A good house is one that provides protection from the weather, comfort and privacy for inhabitants and so on. On the level of specific characteristics, there need be little if anything that is interestingly common to two or more good things. A good knife need not provide protection from the weather, and a good house need not have a sharp edge.[16]

[14] I shall not pursue the common criticism that appeals to nature (as providing direction) involve "the naturalistic fallacy" or illicitly infer an 'ought' from an 'is.' I think the most common general arguments that naturalism is misguided are themselves confused, and, though I agree that no substantive ought-claims can be derived from is-claims that do not themselves presuppose some substantive ought-claim, it is not clear to me that that is what was being attempted in the appeals to nature endorsed by the classical eudaemonists:

> [A]ncient theories are not reductive; in keeping with the way that they do not try to reduce other ethical concepts to those of virtue, they do not try to reduce ethical concepts in general to those that are not ethical....
>
> [T]he notion of nature ... is not a neutral, "brute" fact; it is strongly normative. In defending virtue by showing it to be natural we are not pointing from value to fact, or from evaluative to non-evaluative facts.... For ancient ethics, the facts in question ... are facts which take some finding and the discovery of which involves making evaluative distinctions. (Annas 1993, 135, 137)

[15] See, for example, *Eudemian Ethics* II.1, 1218b 38-1219a 5:

> Let this then be assumed, and also that excellence is the best state or condition or faculty of all things that have a use and a work. This is clear by induction; for in all cases we lay this down: e.g. a garment has an excellence, for it has a work and use, and the best state of the garment is its excellence. Similarly a vessel, house, or anything else has an excellence; therefore so also has the soul, for it has a work.

The word here translated "work" is the same as is rendered "function" in the function argument at *NE* 1097b1 22-32.

[16] I am not yet speaking, nor will I be for some time, of moral goodness.

Yet, the attribution of goodness to (some) knives, houses and innumerable other artifacts is not just a case of homonymy. On a more abstract level, we can see that there is something common to good knives and good houses – that, to some acceptably high degree, they serve the purposes, achieve the goals or fulfill the functions for the sake of which knives or houses are wanted. More generally, a good *x* is one that, to some acceptably high degree, serves the purposes, achieves (or contributes to) the goals or fulfills the functions for the sake of which things of its kind are wanted. The list of characteristics of a good *x* is open-ended because the satisfactoriness of *x* for its purpose or function is a matter of degree and may be increased or improved in ways not previously considered. Additionally, the goodness of *x* is implicitly a comparative matter – the comparison is between the satisfactoriness of *x* for its purpose and the satisfactoriness of some available alternative. This is part of the reason that we cannot replace the clause that demands that a good *x* satisfy its purpose to an "acceptably high degree" with more concrete criteria: what counts as an acceptably high degree depends on the available alternatives.[17]

In speaking of artifacts, I have indifferently referred to a good *x* as answering to the purposes for the sake of which it is wanted and as answering to *its* purposes. And for artifacts, there may be no important difference: they would not exist were they not (believed to be) wanted for certain purposes or to fulfill certain functions.[18] However, the assessment of the goodness of some artifact does not seem to depend essentially upon its

[17] Candles once supplied interior illumination for reading to an acceptably high degree, and so were good to read by, but no longer do so.

[18] I don't think this is the whole story, even for artifacts. The same object may be both a good paperweight and a bad knife. Its being good for the purpose for which it is wanted (being a paperweight) does not make it a good knife. However, it is not important to my current discussion to work out a satisfactory general account of the goodness of artifacts, so I will not pursue it.

being *wanted* – one only needs to know what the purpose or function *is* (together with various facts about the artifact) to judge its goodness. Its being wanted for that purpose may determine *what* its purpose is, but makes no difference to the content of assessments, given that purpose. If you know what its purpose is, you can judge whether or not it is good and how good it is without knowing that it is wanted for that purpose.

This suggests a further possibility. If assessments of goodness do not depend for their content upon the fact that what is assessed is wanted for a given purpose or to fulfill a certain function, but only upon the purpose or function itself, then it may be that the approach can be extended to things that are not or are not known to be wanted – that is, to things with respect to which we do not suppose there to be a conscious designer or intender to impose purposes or functions – provided that there is a purpose or function which itself can be identified. And there are at least *prima facie* plausible cases to be found in the organic world.[19]

- The heart exists in order to circulate blood.
- The eye exists for the sake of sight.
- Sight exists in order to facilitate navigation in a three-dimensional world.
- The acorn exists in order to become a mature oak tree.
- The digestive system exists in order to sort nutrients from wastes.

Examples could be multiplied at length, but I will pause to note a few features that show up in these.

First, though it may be difficult to say what is involved in saying that some

[19] Many of the ancient teleologists, Aristotle included, thought that purpose or function in nature extended beyond biological examples, but we need not follow them that far in order to see the plausibility of attributing it in biological cases.

structure or process exists or occurs for the sake of something else if we are not allowed to appeal to conscious intentions, I take it that these examples really are plausible[20] and that therefore we have reason to see if an account of such purposiveness or end-directedness can be worked out. I will signal that examples of this kind are (we hope) to be explained somehow in terms of their functions or purposes without appealing to conscious design by saying that the relevant explanations are in terms of *natural ends* or *natural functions*.[21] Briefly, we can say that the positing of a natural end or function is an attempt to answer a "what for?" question.

Second, a natural function explanation can be applied to particular organs, such as the heart or the eye, to organ systems, such as the digestive system, and to functions of other organs, such as sight.[22]

Third, natural functions may be hierarchically ordered into those that are more or less proximate. The eye exists for the sake of sight and sight exists to facilitate three-dimensional navigation, or, for a different example, hearts exist to circulate blood and blood circulation exists in order to meet cellular needs. Additionally, structures or processes may be systematically related to the same natural function as the different organs of the digestive system are to the function of sorting nutrients from wastes.

Fourth, and perhaps most important if natural ends are to be applied to give guidance in ethics, the purposes or functions for the sake of which something exists or

[20] One reason it is plausible to explain organs and processes in terms of their purposes or functions is that it is difficult to eliminate such explanations in biology and perhaps even more difficult to do so in medicine. (Try to imagine what medicine would be like if either it had to proceed without use of the concept of health or had to conceive of health without reference to the proper functioning of organisms or their parts.)

[21] I use "natural ends" and "natural functions" interchangeably.

[22] In principle, then, we can ask what is the function of voluntary or intelligent behavior or for the function of moral regulation of behavior.

occurs can be used to assess not only its goodness but the goodness of things that contribute to or interfere with the achievement of the end or the performance of the function. If we know that acorns exist in order to grow into mature oak trees, we can make sensible judgments about what conditions of, e.g., soil, water, sunlight and the prevalence of squirrels, are good or bad for acorns.[23] If we know that hearts exist to circulate blood, we can tell that consumption of fatty foods is bad for hearts and that a leaner diet is better. This point might be turned into a slogan: If you can tell what something is good for, you can tell what is good for it.

If we admit, at least provisionally, that explanations in terms of natural ends may be in order and may help us to grade processes (and supporting or interfering conditions) in terms of the ends they serve, how might this thought to apply to ethics? The basic answer that the classical eudaemonists gave runs in parallel with their accounts of goodness in other cases. We could tell what is good for a human being if we could identify the human function.[24] For the classical eudaemonists, the human function was to be understood as the achievement of *eudaemonia*, often translated as "happiness."[25] To make that more concrete (and leaving aside lots of details), we can say that the function is living a successful life as a mature adult in a social context. That will include possession of the intellectual capacities and traits of character that make possible or contribute to such a life.[26] Since the conditions that make such a life possible or contribute to it are not

[23] Of course, we can extend that to judgments about what is good or bad for the sapling, etc.

[24] *NE* 1097b 22-32.

[25] I think that translation is unfortunate. I discuss why in Chapter Four, "The Structure of Eudaemonism."

[26] I take something like this to be more or less common ground among the classical eudaemonists. For present purposes, I have no comments on the rather puzzling passages in which Aristotle apparently

only external, like conditions of water, soil and sunlight for the acorn, but also depend upon choices, habits and intelligently acquired dispositions, there is room for specifically ethical assessment – for assessment of a person's activities, choices and character insofar as they depend upon voluntary action.

So far, I have tried to portray this kind of account of the human good in a sympathetic light, but plainly, there are questions outstanding both about how we are to understand and identify the human natural function and about the way in which it is relevant to ethics. Both points are crucial, for it might be that the attempt to understand the human good in terms of a human natural function fails in either of two ways. It might turn out, first, that there is no credible way to identify the human function – no way, that is, to identify a function that is both sufficiently determinate to serve in ethical theorizing and which can also lay claim to being objective.[27] Second, even if we can credibly identify something as the human function, it might turn out in any of a number of ways not to be apt for ethical theorizing.[28]

Briefly, I believe that we may be able to give a defensible answer to the first set of questions, but that when we have an account of the human natural end in hand, it will turn out not to provide the kind of guidance we seek for ethics. As a slightly more detailed

elevates the contemplative or theoretical life above a practical life in a social setting as the best kind of life. I am inclined to hope that some kind of reconciliation is possible, but trying to work out such a reconciliation or, alternatively, trying to explain why the apparently discordant praise of the contemplative life is present but unreconciled with other claims, would take us far afield.

[27] It is, of course, no easy task to say what is involved in a method being able to lay claim to being objective, but at least part of what we would want to avoid in such a method can be stated fairly readily: We do *not* want it to be the case that differences in prior moral convictions or intuitions decisively affect the conclusions drawn through the correct employment of the method. More positively, if a method can lay claim to objectivity, we would expect that competent investigators employing it would tend to converge in their conclusions without respect to divergent convictions they brought to the investigation and, moreover, that competent investigators would agree to employ it.

[28] Suppose it turned out that the human natural function was to breathe.

preview of what is to follow, I shall claim that we can understand (and can best understand) natural functions in terms of inclusive fitness but that, if we take such an account *as* an account of the human good, we find first, that it delivers intuitively unacceptable prescriptions in ethics, second, that even if we can identify some more concrete form of individual and social life than contribution to inclusive fitness as the human natural end, that form of life will probably not be accessible to us (and thus will provide *us* with no guidance), third, that more restrictive accounts of the relevance of natural functions to ethics depend for their plausibility upon ethical principles that are not themselves based on natural functions, fourth, that a natural-function-based approach to ethics confuses explanation with justification, and fifth, that it is, despite appearances, ill-suited in general to provide any guidance in cases of ethical perplexity. These are large claims. I shall begin with highlighting certain features of natural function claims, to which any satisfactory analysis should be answerable, and then briefly survey different accounts that have been offered, preparatory to sketching what I take to be the best account, which holds that natural functions are to be understood in terms of inclusive fitness.

1.2 Accounts of Natural Functions

Let us begin by noting certain general features that pertain to claims that something has a function (in the relevant sense).

A function claim is not a claim about how some structure or process actually works or is actually working or what it actually does. The function of a heart is to circulate blood, but, in the first place, it may *malfunction*. If it does not circulate blood or does not do it well, that does not amount to a change in its function. In the second, something with a function may do other things than fulfill that function. A heart also makes noise, but making noise is not its function.

A function claim is not a claim about the statistically typical behavior of a structure or process. If, for example, most people on earth were to simultaneously have heart attacks, that would not mean that hearts had acquired the new function of causing pain and death, but that most hearts were malfunctioning.[29]

A function claim is explanatory – it explains why the structure or process is there, what it is for. Such a claim is not just a gesture in the direction of pointing out the structure's or process's serviceability to some more or less arbitrarily specified end. At least before the development of mechanical clocks, hearts may have been serviceable for the measurement of short intervals of time, but time-keeping is not the function of hearts. Hearts do not exist in order to keep time even if that is one of the things they can be used for.

A function claim is normative, about what a structure or process is *supposed* to do. Hearts are *supposed* to circulate blood and, less directly, to meet cellular needs. Success or failure in doing these things is what enables us to grade hearts as to how well they are functioning. It is this feature that makes it plausible that natural ends could provide guidance in ethics.

How, without calling upon intentions in the mind of a designer, can we make sense of these features of ordinary function claims as they are applied to natural structures

[29] Indeed, something with a natural function may *typically* fail in that function. The natural function of mating calls is to attract a mate but most instances of mating calls fail to do so. (Millikan 1998)

or processes? How can we reasonably and non-arbitrarily attribute functions to them?[30] Or can we do so at all? Might it be that the functions we attribute just reflect our purposes and interests or, perhaps, our laziness with respect to working out a causal explanation for the structures and processes in which we are interested? Setting aside access to the intentions of a designer, what options are available to account for these features of natural functions?

1.21 Value-Based Accounts

Since he is often thought to be both the paradigmatic teleologist and the paradigmatic eudaemonist in the philosophical tradition, it is useful to begin by looking at Aristotle's use of teleological thought. Fred Miller helpfully distinguishes four features of teleological explanations to be found in Aristotle's work, "(*a*) involving a potential for form which cannot be reduced to the powers of the material elements; (*b*) happening for the sake of something good; (*c*) having intrinsic causes and hence not being mere chance outcomes; and (*d*) involving an inherent self-regulating principle,"[31] and points out that which is taken to be the most fundamental makes a difference to the interpretation of Aristotle. Since my concerns here are not primarily exegetical, I will focus mainly upon (*b*) and (*d*).

Miller suggests that there is some advantage to taking (*d*) to be the most fundamental feature since the other features can themselves be explained in terms of the

[30] If we had insights into a designer's intentions, that would be a very good way of reasonably and non-arbitrarily attributing functions to the products of her design, but I do not suppose we have such insights, at least none upon which competent investigators can be brought to agree.

[31] Miller 1995, 340.

presence of an internal directive principle. I can accept his arguments with respect to the explanatory power of an internal directive principle for (*a*), irreducible potential for form, and (*c*), intrinsic causation, but I think that it does not adequately capture (*b*), explanations in terms of what is good – hereafter, value-based explanations.

To the contrary, internal directive principles are themselves best understood in terms of value-based explanations, and therefore, it is value-based explanations that should be regarded as the most fundamental. Consider the argument from internal directive principles to value-based explanations. A plausible case of an internal directive principle would be the genetic program, encoded in an organism's DNA, that guides the organism's development to maturity. In terms of that encoded program, we can say why a given feature of an organism is good for it – that is, why it contributes to the organism's having the kind of life specified in the genetic program. We can also identify certain things that could go wrong with the normal developmental program. But this is not a sufficient account of the goodness of such species-typical features for it provides no reason to suppose that the mature form specified by the genetic program is itself good or part of the organism's good.

This can be made clearer by noticing two different ways in which features of an organism may be defective. There are, first, what may be called developmental defects. Some external cause may interfere with the normal developmental program or some necessary supporting condition of normal development may be absent. A mother may drink too much during pregnancy, and the alcohol interferes with the infant's normal development. This would be a case of an interfering external cause. Or, a mother may be malnourished during pregnancy and the infant's development is thereby stunted. This would be a case of the absence of a normal supporting condition. However, there is a different way in which features of an organism may be defective. These may be called original defects. Suppose that there is a properly genetic defect in the developing organism, so that the actual genetic program at work in the development of an infant does not encode for the development of arms. Then, there need be neither the influence of external causes interfering with development nor the absence of supporting conditions that are normally present. However benign the external causes or the normal supporting conditions, the infant will not develop arms. Its defect is not that it fails, in one way or another, to realize the developmental pattern encoded in its genetic program. It *does* realize that pattern, but the pattern that it realizes is not good for it (or not as good as the more common pattern realized in other members of its species).

It is value-based explanations, then, that are more fundamental than those in terms of internal directive principles[32] because, though internal directive principles can account for what goes wrong in developmental defects, value-based explanations can account also for what goes wrong in cases of original defects. Accordingly, the best option for an account of natural functions in an Aristotelian framework is value-based. If we are looking for something else, we will have to go significantly beyond Aristotle.

Now, I do not think that, in general, I need to resist appeals to value-based explanations to account for natural functions.[33] However, there is a reason such an account is not apt for our current purposes. Since our concern is to identify the human

[32] I do not mean to be claiming that this was Aristotle's position. I do not know if he considered the question or, if he did, what he would have thought about which feature of teleological explanations was most fundamental. I am only claiming that it is value-based explanations that *we* should take as most fundamental in understanding his position.

[33] For a contemporary defense of a value-based account of natural functions, see Bedau 1998.

natural function (or human natural functions) in order to discover what contributes to or constitutes living well, we cannot rely upon an account of natural functions in terms of which those functions themselves will be explained via the well-being of the organism.

1.22 Cummins' Causal Role Account

If we set aside value-based accounts of natural functions, there are two further major contenders. One of these is the etiological account that will be explained and defended at some length below. The other, which I wish to examine now, is the causal-role account developed by Robert Cummins.

Fortunately, the examination can be relatively brief, and a detailed presentation of Cummins' account is not necessary. This is not because Cummins has failed to elucidate a scientifically useful concept. Rather, it is because the concept he analyzes, however useful for some purposes, is not suited to play the role teleologists require of natural functions. In short, though the concept he elucidates may have a legitimate role in inquiry, it is not the same as the one to which teleologists appeal and therefore is not really competitive with it.[34]

Briefly, Cummins holds that the analysis of functional explanations has been derailed by the assumption, which he calls (A)[35]: "The point of functional characterization in science is to explain the presence of the item (organ, mechanism, process, or whatever)

[34] We might think it competitive if it turned out to be the only intelligible conception of function, but it is not. The next section will show that an analysis better suited to the teleologists' requirements is available.

[35] There is also an assumption (B) that Cummins finds problematic; however, it is, according to him, problematic primarily in the way it is interpreted when conjoined with (A). (1998, especially 169, 179-184)

that is functionally characterized."[36] Cummins thinks we will do better if we reject (A):

> To attempt to explain [for example] the heart's presence in vertebrates by appealing to its function in vertebrates is to attempt to explain the occurrence of hearts in vertebrates by appealing to factors which are causally irrelevant to its presence in vertebrates. This fact has given "functional explanation" a bad name. But it is (A) that deserves the blame. Once we see (A) as an undefended philosophical hypothesis about how to construe functional explanations rather than as a statement of the philosophical problem, the correct alternative is obvious: what we can and do explain by appeal to what something does is the behavior of a containing system. (Cummins 1998, 176)

For Cummins, the functional characterization of some item does not provide an explanation for the presence of that item but rather refers to its causal role in producing some effect in a system of which it is part. Plainly, if an account of this kind is accepted as the one relevant to natural functions, they will not only not be explanatory of the presence of items functionally characterized but will also not account for the other common features of function claims noted above. That is, there will be no account of the divergence of function claims from the actual or statistically typical behavior of the item nor will the normativity of function claims be captured. Perhaps something like that will have to be accepted in the end, but if it must be, that will amount to giving up on the prospects for appealing to natural functions in ethical theorizing. It is worth considering

[36] Cummins 1998, 169.

an alternative that appears better suited to the teleologists' purposes.

1.23 The Inclusive Fitness Account

The kind of account of natural functions I shall try to sketch draws on the work of many recent thinkers in both biology and philosophy who have worked out slightly differing versions of what has come to be called the *etiological account* of natural functions.[37] It will not be necessary, for my purposes, to deal with the sometimes subtle distinctions that are made between these different versions.[38] I will confine myself to outlining the general picture that they share[39] and will also try to provide an answer to the argument that functional explanations are superfluous, that they substitute relatively easy armchair theorizing for more fundamental causal accounts.

[37] The *locus classicus* for this kind of account is Larry Wright's 1973 article, "Functions." (1998) Its most sophisticated version, in my judgment, can be found in the work of Ruth Millikan. (1984; 1998) Also important, for its response to certain features of Millikan's account, is Karen Neander's "Functions as Selected Effects: The Conceptual Analyst's Defense." (1998)

[38] An excellent collection of articles exploring versions of and alternatives to the etiological account is *Nature's purposes: analyses of function and design in biology* (Allen, Bekoff, and Lauder 1998).

[39] Some friends of natural functions have proposed less restrictive accounts than those to which I shall be referring. On one hand, it is easier to satisfy the conditions of less restrictive accounts, but, on the other, it is harder to see why satisfaction of those conditions might be thought to be ethically relevant.

For example, Eric Mack says that "[t]he nature of the function of something can be determined by the requirement which accounts for the existence of that thing" and elaborates that the existence of the function involves the existence of "some need or requirement which explains (plays a role in explaining) the existence of some thing (object, activity, process, etc.)." (1971, 735) There is a problem, however, in understanding this. What are we to make of something being a need or requirement? If to be a need or requirement is just to be something for which there is some causally necessary condition, then there are natural ends or functions everywhere. The need or requirement to destroy the World Trade Center explains the trajectory of airliners headed towards it. The natural function of obesity is to discourage exercise that would result in weight-loss. The natural function of the secret police is to suppress dissent that might overturn a totalitarian state. And so on. Someone may of course stipulatively use "natural ends" in that way, but it doesn't look very interesting ethically. It only becomes interesting if one forgets all the things the definition lets "natural ends" apply to and focuses upon a subset of them that are found ethically salient for reasons other than that they are natural ends.

Alternatively – which is what I suspect Mack had in mind – the appeal might be to some notion of genuine or objective needs that form only a subset of causally necessary conditions. Unfortunately, the notion of such objective needs is as much in need of analysis as the original idea of natural functions. So understood, Mack's account does not amount to a solution so much as to a relocation of the problem.

We can begin by considering how biologists think about adaptations. In order to do this, some terminological distinctions are needed. We need to understand what an adaptation is and how it is related to and distinct from a feature which is adaptive. What an adaptation is can best be understood in terms of inclusive fitness. The inclusive fitness of an organism is relative to an environment which may and typically does include other organisms. That organism will possess certain heritable traits. The inclusive fitness of an organism is a measure of the contribution it makes, *because* of those heritable traits (not just accidentally), to the presence of other organisms carrying those same heritable traits in subsequent generations.[40] That contribution will typically occur through the organism's own successful reproduction, through what it does for close relatives (who will normally be carriers of at least some of the same heritable traits), through reciprocal altruism,[41] or some combination of these.[42]

Given this, a heritable trait is *adaptive* if it contributes to inclusive fitness. That is, an organism possessing it is more inclusively fit than an otherwise similar organism (living at the same time and in the same environment) lacking it or possessing some (actual) alternative to it.

If a trait is adaptive and if the environmental conditions under which it is adaptive

[40] "Inclusive fitness is calculated from an individual's own reproductive success plus his *effects* on the reproductive success of his relatives, each one weighed by the appropriate coefficient of relatedness." (Dawkins 1982, 186)

[41] One organism may act in a way that benefits another that is not closely related because the other can be expected to reciprocate. At least some cases of symbiosis can probably be explained in this way.

[42] If the relevant information were available, then, for comparative purposes (this variant is more inclusively fit than that), in principle, an index number representing an organism's inclusive fitness could be assigned by determining how many other organisms carrying the same heritable traits there can be expected to be (where the expectation is based on the heritable traits) in subsequent generations because of that original organism. Thus, for example, if there were a population of genetically identical organisms that doubled in size in each generation, then the inclusive fitness of each member of that population could be represented by the index number, 2.

remain stable, it can be expected that the trait will spread through the population; its carriers will reproduce more successfully and will be less likely to be eliminated under adverse conditions than non-carriers. This is to say that there will be selective pressure in favor of the trait.

An *adaptation* is a trait which is present in a population of organisms because there has been, at some time and in some environment, selective pressure among its ancestors for that trait.[43] An adaptation may be more or less complex and the clearest examples will be of complex traits that must have been shaped out of multiple mutations.

Clearly, if these definitions are accepted, it is not necessarily the case that an adaptive trait is an adaptation. First, there may be adaptive traits that are not heritable. A hunter's skill is not genetically transmitted to offspring. Second, there may be traits exhibited in a population which are both adaptive and heritable but the explanation of which does not include selective pressure in their favor. This would be the case for any new mutation that is adaptive and would also be the case for any adaptive traits that are in some way byproducts of other processes.[44] It is also not necessarily the case that an adaptation must be adaptive. It must have *been* adaptive under the circumstances in which it evolved but those circumstances may be quite different from what the organism carrying it faces now.[45] There is no serious doubt, for example, that the human appendix

[43] I shall not place much emphasis on what is or is not to count as a trait. Since we are speaking of heritable traits, it will have to be the case that any adaptation referred to has some realization in the body or brain of the organism in question. However, our evidence for the existence of an adaptation will often be largely or entirely behavioral. We may have only indirect arguments that the trait is realized in or supported by some inherited structural feature of the organism.

[44] See Gould 1991. To complicate the story slightly, it should be noted that a trait may *emerge* as a by-product of other processes, but be preserved because it confers adaptive advantages. See Dennett 1995, 238-251.

[45] Steven Pinker puts this nicely in an application to human psychology (1997, 207-208):

is an adaptation, though it no longer makes any positive contribution to our reproductive success.

To connect this with the terminology of natural functions, we can say that an adaptation, whether it be an organ, an organ system, a process or a behavior, is a trait that has a natural function and that it has such a function just when there is an explanation based in natural selection – that is, one ultimately in terms of contributions to the organism's inclusive fitness – for its presence or features.[46]

An etiological account of this sort, which interprets talk of natural functions in terms of the causal history that gave rise to that which is said to have the function, seems able to capture at least three of the four features mentioned earlier of the way that we talk about functions. It allows a distinction between the actual or statistically typical working of a structure or process and its function and allows us to understand why the attribution of a function is normative rather than merely descriptive. Some, however, have thought

> … [W]hat about the Darwinian imperative to survive and reproduce? As far as day-to-day behavior is concerned, there is no such imperative. People watch pornography when they could be seeking a mate, forgo food to buy heroin, sell their blood to buy movie tickets (in India), postpone childbearing to climb the corporate ladder, and eat themselves into an early grave. Human vice is proof that biological adaptation is, speaking literally, a thing of the past. Our minds are adapted to the small foraging bands in which our family spent ninety-nine percent of its existence, not to the topsy-turvy contingencies we have created since the agricultural and industrial revolutions. Before there was photography, it was adaptive to receive visual images of attractive members of the opposite sex, because those images arose only from light reflecting off fertile bodies. Before opiates came in syringes, they were synthesized in the brain as natural analgesics. Before there were movies, it was adaptive to witness people's emotional struggles, because the only struggles you could witness were among people you had to psych out every day. Before there was contraception, children were unpostponable, and status and wealth could be converted into more children and healthier ones. Before there was a sugar bowl, salt shaker, and butter dish on every table, and when lean years were never far away, one could never get too much sweet, salty, and fatty food. People do not divine what is adaptive for them or their genes; their genes give them thoughts and feelings that were adaptive in the environment in which the genes were selected.

[46] Explanations of this sort can be extended to cases in which there is not literal biological reproduction or inheritance, as in the first two chapters of Millikan 1984. Less systematically developed, Richard Dawkins' notion of "memes" is also of this type. (1976)

that an account like this does not successfully capture the feature that natural functions must be explanatory. This deserves further attention.

In *Philosophy of Social Science*, Michael Root has recently surveyed and endorsed some general criticisms of functional explanations. Though his concern is primarily with the employment of functional explanations within the social sciences, if his general criticisms are correct, they would also (as he recognizes) apply to functional explanations in evolutionary biology. Those of his criticisms which are relevant to the issue of the explanatory power of natural functions[47] can be summarized as follows. First, they do not solve what he terms "the selection problem." Second, they do not explain why a trait, T, rather than a functionally equivalent trait, T′, prevails in a population.

Root's first claim is that functional explanations (where they are not mediated by the deliberate choices or intentions of some designer) fail to adequately address what he terms the selection problem – "why of all the solutions that might have been selected [to a hypothetical design problem] T was selected." (1993, 83)

> An answer to the selection question must include a description of how the trait is transmitted from some members of the group to others. The biologist explains how, first by theorizing that the trait is the effect of a gene and next by describing the process by which genes are inherited. However, this answer to the selection question replaces a functional

[47] Root also objects that explanations in terms of functions are empirically empty because one can always postulate an adaptive problem in a hypothetical ancestral environment for which the trait in question appears to provide a solution. The answer to that is surely, as Elliott Sober says, "[i]f optimality explanations are too easy to invent, let's make the problem harder." We impose empirical constraints on what counts as a plausible description of the ancestral environment, what counts as a success for the proposed explanation (does it, for example, *predict* what is observed rather than simply provide a just-so story?), and so on. (1993, 134)

> explanation with a causal one....
>
> Functionalism faces a dilemma.... If it doesn't offer an answer to the selection question, then it doesn't explain the presence of the trait; but if it does provide such an answer, then the answer, rather than the functional fact, explains the presence of the trait. Add the causal facts needed to explain the selection of the trait, and the fact that the trait is functional is trimmed from the explanation; for the causal rather than the functional fact now explains the presence of the trait.[48]

In a closely related criticism, Root maintains that functional theories in evolutionary biology do not deal adequately with the problem of functional equivalents. For any given trait, T, for which a case can be made that it is adaptive with respect to some problem faced by an organism, there is some other possible trait, T′, which would be equally adaptive with respect to that problem. However, the only explanation that the biologist has to offer as to why T appears in the population rather than T′ is that genes for T and not for T′ appeared among the organism's ancestors. Once again, a functional explanation is replaced by a causal explanation. (1993, 86-88)

I am addressing both of these together because it appears that they both rest on the same mistake – roughly, upon the assumption that causal explanations are invariably competitive with functional explanations. A useful way of showing that they are not is in terms of the model of fitness landscapes.[49]

[48] Root 1993, 86.

[49] For some discussion and further references, see Dennett 1995, 190ff. Further useful discussion may be found in Nagel 1968, 80-96.

One imagines the ancestors of an organism at some time in their history located upon an abstract landscape, a fitness landscape, where ascents represent increases and descents represent decreases in inclusive fitness. Where exactly the organisms ascend (or not) depends on causal factors – what adaptively favorable mutations occur. These will be hills (or mountains) on the adaptive landscape. So, in one sense, there is a causal explanation for every adaptively favorable mutation that spreads through a population. However, if there had been a mutation producing a different but functionally equivalent trait present (then and there) on the fitness landscape, that trait would have spread through the population instead. The functional explanation focuses on what is common to the two or more different causal stories, namely, that they both contribute (or would have contributed, had they occurred) to inclusive fitness.

The point can be illustrated with a non-biological example: What explains the stable and non-interfering orbits of planets in the solar system? Why are there not many bodies in unstable orbits? How, in short, were the stable orbits selected (the selection problem) and why *these* stable orbits (the problem of functional equivalents)? Now, there is, of course, for each planet, a causal account that explains why it is in the particular stable orbit that it is. Apparently, what Root would recommend is that we be satisfied with taking the conjunction of the separate causal accounts as the explanation for the overall order of the solar system.

But there's a simpler explanation, obvious once it has been suggested, that operates on a different level of generality. To wit: the planets we see in stable orbits are *survivors*. There may have been any number of bodies in the solar system in unstable orbits. But, given time (and in the absence of any regular or large-scale influx of bodies in unstable orbits from outside the system), they either fell into the sun, achieved escape velocity from the system, or collided with something else, thus producing one or more new bodies which then, if the collision did not result in its product(s) achieving stable orbit, either fell into the sun, achieved escape velocity, or collided with something else ... and so on.

Eventually, any solar system, in whatever state of order or disorder it may have begun, can be expected to be either empty or to be occupied primarily by bodies in stable orbits.[50] That is, though there is a detailed causal explanation for each stable orbit, there is also a functional explanation for the fact that almost all the orbits are stable. Since the explanations are not competitive with one another, the correctness and relevance of the detailed causal account does not, as Root presupposes, necessarily undermine claims that a functional account is *also* correct and relevant.[51, 52]

[50] I am still assuming the absence of any large-scale or regular influx from outside the system.

[51] The qualification, "not ... necessarily," is important. Some causal accounts would undermine some functional explanations. If, for example, we had reason to think that basic laws of celestial mechanics ruled out unstable orbits, then the functional explanation would be an unnecessary fifth wheel.

[52] It might be wondered whether, if we can find functional explanations appropriate for the explanation of non-biological systems, there remains any interesting normative punch to the claim that natural functions are important to understanding living organisms. Are we saying any more when we say, e.g., that the heart is *supposed to* pump blood than when we say that the planets are "supposed to" be in stable orbits?

I think we can identify a difference between the two cases. The most promising suggestion in this direction that I know comes from Nozick. His suggestion is that, for full-fledged natural functions, we need a two-level etiological account: "Z is a function of X when Z is a consequence (effect, result, property) of X, *and* X's producing Z is itself the goal-state of some homeostatic mechanism [or process, such as conscious design or natural selection] ... and X was produced or is maintained by this homeostatic mechanism M (through its pursuit of the goal: X's producing Z)." (1993, 118) In other words, some organ or process will have a function if it is "designed" to have that function and the "designing" itself can be understood as a homeostatic process. See also Nagel 1968.

If, however, I am not correct on this, there is still a larger point to be made. If we cannot identify anything interestingly normative in any functional explanations, that is compatible with, and in fact would lend support to, what I aim to show in this chapter, namely, that natural functions do not provide the right kind of guidance for purposes of ethical theory.

1.3 Are Natural Functions Fruitful for Ethics?

I have sketched the kind of account of natural functions that can be developed in terms of inclusive fitness at some length and responded to what I think are the most important objections in order to make it clear that, why and how I take talk of natural functions to be scientifically respectable. So far, it is still an open question whether natural functions can be used to support ethical conclusions in the way the classical teleologists thought (though with an empirically better informed account of what natural functions are).

Before proceeding, however, it may be asked whether it is fair to natural law thinkers to attribute to them or to interpret their positions in terms of an account of natural functions that was not available to many of them. I think there are grounds of two sorts for saying that it is. First, there does not seem to be anything better (or as good) available. If *anything* can play the role natural law thinkers have reserved for natural functions, then it is the kind of functions that can be identified by the etiological account. Second, at least among modern thinkers working within the tradition, the need for some such account seems widely appreciated. Terence Irwin cites Wright's version of the etiological account.[53] So do Richard Sorabji[54] and James Wallace.[55] Eric Mack, apparently independently, works out a version of an etiological account,[56] and Roderick Long adapts Wright's account.[57] Fred Miller situates natural teleology within evolutionary

[53] Irwin 1980, 51.

[54] Sorabji 1980, 160. Sorabji's discussion, unlike the others cited, is not especially concerned with the relevance of natural teleology to ethics.

[55] Wallace 1978, 23.

[56] Mack 1971, 735.

[57] Long 1993, 11-19.

theory.[58] Not all are as clear as might be wished upon the matter, but the convergence upon something like an etiological account is impressive testimony both that something of the sort is needed and that nothing better is available.

So, given an etiological account of natural functions, what can we say about whether there is something that can be identified as the *human* natural function? (And if there is, what is it?)

There appear to be three options for theorists. First, it may be that inclusive fitness itself is the relevant natural function. Second, it may be that the human natural function involves living a more concretely specifiable individual and social life. Third, it may be that there is no single natural function in terms of which to guide one's life, but that natural functions may be used to identify goods that arguably should have a place in a good life. I think the first of these is the most defensible account of the human function but the least appealing morally. Immediately below, I shall undertake to defend that interpretation and to show that if contribution to inclusive fitness is taken to be the ethically relevant human function, it leads to counter-intuitive consequences. Then, I will address the other two options and also develop two more general objections which, first, do not require appeal to moral intuitions, and second, apply to all of the options.

1.31 Inclusive Fitness as the Directly Relevant Natural Function

Biologists do not often, unless they are ecologists, speak about the functions of whole organisms;[59] they confine themselves to discussing the functions of parts, processes

[58] Miller 1995, esp. 344f.

[59] When functions are attributed by ecologists to organisms or species, it is typically to their functions within ecosystems. Such attributions plainly cannot be explained in terms of the contribution of

and behaviors. But there seems no reason that the kind of account indicated above cannot be applied to understand what the function of an organism is. We can ask whether there is any over-arching function of the organism, or at least of such of its behaviors as are directly or indirectly under voluntary control, that promote the organism's inclusive fitness. I think, however, that we will not find a great deal of illumination from that direction. The only natural function we will be able to find for human beings will be just the promotion of their own inclusive fitness.

The crucial reason for this is that, in complicated organisms like ourselves, natural functions are highly modular. We can specify the functions of the heart, the liver, the visual processing system, the language centers in the brain and so on in quite a bit of detail. We can explain why organisms with those phenotypic features[60] would be more likely to successfully reproduce than similar organisms with slightly different designs.

But the natural functions of the parts or modules don't add up in any very illuminating way to a natural function or end for the whole organism. About all that can be said, if one wants to talk about the organism's natural function, is that it must be (or its ancestors must have been) good at reproducing.[61] We can say that the digestive system must be good at sorting nutrients from waste in dietary intake. We can say that the visual

those organisms to their own inclusive fitness, but perhaps a more abstract etiological theory, such as Millikan's, can accommodate them. Alternatively, it may be that some different analysis, such as Cummins', is needed.

[60] Briefly, the *genotype* is the heritable genetic "recipe" for building an organism. The *phenotype* is the way the genotype is expressed. Since the genotype will be expressed differently under different environmental conditions, genotype underdetermines phenotype. (Height, for example, is certainly affected by the genotype, but increase in average height over the past several decades is due to nutritional rather than to genetic changes.)

[61] Technically, it doesn't have to be good at reproducing *itself*; it just has to be good at contributing to the existence of organisms in the next or subsequent generations that carry the same genes and it has to be *non-accidentally* good at doing so. Its phenotype has to typically result from its genotype and typically result in such a contribution to reproduction.

system must be useful for navigating a three-dimensional world. But what must the digestive system and the visual system *together* be good at? Only one thing: contributing to reproduction.[62]

In itself, this may seem somewhat disturbing. Contributing to reproduction or, more accurately, to inclusive fitness doesn't seem as promising for ethics as "activity of soul in conformity with excellence.... in a complete life"[63] or the Stoic ideal which identifies happiness (or *eudaemonia)*, the life according to nature, and virtue.[64] Be that as it may, it is worth examining further. I shall begin by considering three sorts of cases in which it appears that accepting inclusive fitness as the (ethically relevant) human function leads to counter-intuitive consequences.

First, if inclusive fitness is the human function, there appear to be cases in which it provides no guidance at all. Some people may not be in a position to contribute to their own inclusive fitness – for example, some of the elderly who are themselves past reproductive age and have no living relatives. Such people may still enter into decent, respectful, honorable and humane relations either among themselves or with others. Of course, they also may not, but it is, I submit, counter-intuitive to hold, as one would have to if inclusive fitness is the touchstone for ethics, that they have no moral reason for preferring decent, respectful, honorable and humane relations to their alternatives.

Second, there appear to be cases in which the appeal to inclusive fitness gives the

[62] I take no position on whether the contribution to inclusive fitness may or must be explained (for some traits that it is plausible to regard as morally relevant) through some kind of group selection. See Richards 1995, 259ff.

[63] *NE* 1098a1 16-18.

[64] "Thus, on the Stoic theory there is for each of us some single end that it is appropriate for us to refer everything we do in life to, and this end is identical, first, with our own 'happiness' ... second, with our living in agreement with nature, and third, with our living virtuously." (Cooper 1996, 261)

wrong answer to ethical questions. For example, it would justify rape by conquering soldiers. The soldiers are young and healthy, but may die at any time in battle. In their circumstances, rape gives them the best chance at passing on their genes.

Third, the appeal to inclusive fitness gives the wrong kind of answer in another respect. It may give the right answer for the wrong reason. It may be true that some wrong action does not contribute to or interferes with achieving or promoting one's inclusive fitness. But that may be a complete irrelevance. Consider the case of Adolf Hitler, whose actions resulted, among other things, in the deaths of several million Jews. It is plausible – and I invite you to suppose that it is true – that his actions did not promote his inclusive fitness. Had he not been so obsessed with the Jews, he could have sired more offspring.[65] So, in this case, the appeal to inclusive fitness gives the right answer about what he should have done. He should not have made the plans, taken the actions, or cultivated the obsessive hatred that led to those deaths. But surely, though this is the right answer, it is reached for the wrong reason. What Hitler did was wrong, not primarily (if at all) because of the way those actions interfered with his reproductive success, but because those actions trampled on the interests and violated the rights of his victims. Hitler's moral failure was *not* that his obsession kept him from having as many children as he otherwise could have.

The direct appeal to inclusive fitness as the human natural function, then, yields counter-intuitive results in at least three ways. It sometimes tells us that moral considerations are irrelevant where we are confident that they are not, it sometimes tells us that acts that we are confident are morally wrong are in fact morally right or at least

[65] So far as I know, he had none.

permissible, and it sometimes gives us an account of why certain actions are morally objectionable that we are confident is mistaken. The point can be put more strongly in the reverse direction: If our intuitive responses to these cases are correct and if it is also correct that promotion of inclusive fitness is the human natural function, then, for the first and third cases, our action ought to be guided by something *other than* our natural function, and, for the second case, we ought to act *against* our natural function.

1.32 Beyond Intuition

However, this is less than decisive in at least two respects. In the first place, appeals to intuition are not decisive. I think that widespread moral conviction, especially when shared by persons who otherwise differ in moral theory, should be taken very seriously, but there is still the possibility of a shared mistake. In the second place, the arguments that an appeal to a natural end leads to counter-intuitive conclusions are predicated upon the (argued) assumption that the only natural function for human beings (as distinct from particular organ systems, etc. *of* human beings) that we can identify (consistently with a scientifically respectable account of natural functions) is inclusive fitness. However, there are, as mentioned above, available two other understandings of the relevance of natural ends to ethics. Accordingly, I shall first address each of those and then develop two more general objections to the appeal to natural teleology in ethics that do not depend either upon assuming the correctness of moral intuition or upon the assumption that the relevant natural function must be the direct promotion of inclusive fitness.

1.321 A Concrete Life as the Human Function?

One of the two alternative readings of the relevance of natural ends to ethics is that human beings have it as their natural function to live some concretely specifiable individual and social life[66] – one, that is, which is *more* concretely specified than just the direct promotion of inclusive fitness. Let us suppose we can credibly identify some kind of individual and social life to which we are naturally adapted. If so, it is almost certainly not accessible to most of us. The evolutionary processes in terms of which we can understand natural functions are very slow, especially in comparison with the processes of cultural change operative in human societies throughout recorded history. Human biology has been essentially fixed for approximately two hundred thousand years. Therefore, we should expect that if some more concrete end than success at contributing to reproduction – some kind of individual and social life – is or is part of our natural end, then the individual and social life in question would be that which was characteristic of our ancestors at the time that our biology was fixed. Since those ancestors (and indeed, probably the majority of biologically modern humans throughout our species' tenure on this planet) lived in hunter-gatherer bands, we should expect that we are naturally adapted to life in hunter-gatherer bands. But if this is so, the life which it is our natural function to live is one that, for the vast majority of us, is not an available option. If everyone were to attempt to live that kind of life, most of us could not survive. At best, such a natural function could provide no guidance to most of *us*.[67]

[66] Such a natural function, of course, would itself have to be explicable in terms of its contribution to inclusive fitness.

[67] This is powerful, I think, but not quite decisive. It might be held that there is some appropriate

1.322 Multiple Natural Ends

Another way of interpreting the relevance of natural ends to ethics would involve abandoning the assumption that there is some *single* natural end or function in terms of which to guide one's life. Since the classical teleologically-oriented eudaemonists shared the assumption of a single end or function, it is appropriate in discussing their positions, but one might take a less global view of what natural functions can be expected to do. We could instead identify various functional characteristics, natural desires perhaps, and claim that their objects are goods for us.[68] To some extent, my criticisms of appealing to a natural function to pick out an over-arching end will apply to this more modest proposal as well. But even where it does not, there will be, on this proposal, a need for some kind of ordering principle or combinatorial function to deal with cases in which goods cannot be jointly realized and, where necessary, to direct us to appropriately subordinate one

level of abstraction at which a human natural function, identified neither with the direct promotion of inclusive fitness nor with the concrete social arrangements prevalent when our biology was fixed, can be picked out. To say the least, accomplishing that – including the provision of a credible case that such a life *is* (or is part of) our natural function – will be a difficult task. Even if it can be pulled off, however, I think it would fail to meet the more general objections yet to be developed.

[68] See, e.g., Arnhart 1998. It should be noted that there may well be cognitive, motivational and behavioral adaptations that bear on our susceptibility to being moved by moral considerations. There is, for example, considerable evidence that we are much better at detecting cheaters on social rules (those who collect benefits without paying the rule-assigned costs) than upon comparable problems not involving such detection. (Cosmides and Tooby 1992) That, of course, would be important for prospective cheaters to take into account. Unfortunately, there's another side to it: better detection breeds better cheaters. There's not likely to be a persuasive argument that *all* of us have a natural function of avoiding cheating. Quoting Kim Sterelny:

> The psychological disunity of mankind is no idle possibility. One plausible hypothesis predicts disunity. [Robert Frank's ideas about the emotions as commitment devices] can ... be used to make a point about the diversity of human cognitive design. For if he is right, then having emotions incurs a cost: really cooperating, rather than pretending to, and cheating. So, we should expect an evolutionary arms race between the emotional and emotion mimics, who try to parasitise, getting benefits but not paying costs. We should expect long term survival of both mimic and model.... So the commitment problem suggests that selection might maintain a diversity of human psychologies. There is no single best design for solving it. (1995, 378f.)

good to another, and that principle, *ex hypothesi*, is not given to us by any natural end.

1.323 Two More General Problems

There are two further ways in which natural ends give the wrong kind of answer for ethics, which apply independently of the particular readings given of their relevance to ethics. First, they provide explanations rather than justifications. Second, they do not in fact provide the kind of guidance we seek from an ethical theory. Each point bears some elaboration.

1.3231 Confusing Explanation with Justification

We may explain why we have certain tendencies, desires, behavioral patterns and the like by their contribution to inclusive fitness. That does not tell us either *why* we should have them or *that* we should. An explanation in the present context, where we are speaking of actions or dispositions, says why an action occurs or why a motivation or behavioral tendency is present. A justification would say why it ought to occur or why it is better that the motivation or tendency be present. A justification can be (or be part of) an explanation: you can explain why a person did something by saying what justified her in doing it. However, there are explanations for things that are not justified – child abuse, for example, is not inexplicable just because it is not justified.

If, to give an example, there is an evolutionary explanation for xenophobia,[69] for the tendency to divide people into "us" who count, who are important, and "them" who

[69] There probably is such an explanation. It would have to do with the way in which xenophobic tendencies promoted the inclusive fitness of members of small and generally closely-related members of hunter-gatherer bands in competition with other bands of hunter-gatherers who were much more distantly related.

do not – who are at best neutral, but more commonly enemies or resources to be exploited – that tells us only that such a behavioral pattern helped our ancestors reproduce, not that it was admirable and not that we have any reason to emulate them now.

This is connected to the (much) earlier point that a feature of an organism may be an adaptation – that is, it may have been shaped out of mutations in response to a problem that organisms of that type faced in their ancestral environment – without being presently adaptive. Even if we overlook or set aside the question why or whether we are justified in serving some particular natural function, an adaptation may, under current conditions, confer no advantages or be a disadvantage. Even in terms of what made the trait functional, its contribution to inclusive fitness, there may be no reason for continuing to have the trait, and, if it manifests itself in the form of behavioral tendencies or motivations that we are capable of resisting or suppressing, there may be reason for resisting or suppressing them. Our natural xenophobia may, under present conditions, increase the likelihood of war on a scale that may prove destructive both to "us" and to "them." It seems excessively respectful of nature's handiwork to suppose that, even in such a case, we ought to give our natural xenophobia free rein rather than resisting and suppressing it.

1.3232 Failure to Provide Guidance

There is also a further and deeper problem with the appeal to natural ends as a touchstone in ethics. Appearances to the contrary perhaps, natural ends, assuming they can be identified, do not in fact provide guidance if we are in doubt about what to do. For there are, in general, two possibilities. A natural end may inexorably control behavior, either directly or by way of some irresistible motivation. But if that is so, no question of choice, of what one should do, arises. If there are such natural ends, agents *will* act accordingly. They will face no moral choice about whether or not to serve such ends.

Alternatively, the natural end in question may not ineluctably control behavior. Though we may suppose that the end can somehow be identified, it leaves the agent at most with only tendencies or motivations that can be resisted. It may even be that she will not be left with so much as tendencies or motivations, just the possibility of reasoning about what would serve or best serve her natural end. She can still ask whether to go along or resist. The natural end provides tendencies and motivations. Reasoning about it can tell her both what actions would promote it and perhaps apprise her of the frustrations she is likely to experience if she does not promote it. But nothing about it, simply insofar as it is a natural end, tells her that she *ought* to promote it.

Once the question is clearly posed and clearly understood, the agent may, of course, decide that service to the natural end is her best option. But that decision will not be justified solely in terms of the fact that the selected act or plan serves the natural end. If the agent has a reason for selecting it, *that* reason must come from somewhere else. Knowledge about the natural end and the motivational and behavioral tendencies to which it gives rise at most provides information to be taken into account in deciding what to do. The decision itself must be justified in some other way.

1.4 Summary

This has been lengthy and summary may be useful. The classical eudaemonists typically defended their conceptions of the appropriate end in terms of which to guide one's life by an appeal to natural teleology. Though some have been suspicious that natural teleology is a relic of an out-dated and no longer defensible world-view, it is possible to make scientifically respectable sense of the notion. In evolutionary terms – specifically in terms of inclusive fitness – we can understand what it means to speak of natural ends or functions. However, once we import that understanding into the consideration of specifically ethical arguments, we find, first, that the direct appeal to inclusive fitness leads to counter-intuitive ethical conclusions, second, that insofar as natural ends can be understood to specify some concrete individual and social life, what is proposed as an ethical ideal is a way of life not open to most of us, and third, that less global appeals to natural ends will be incomplete unless supplemented by some principle the rationale for which cannot itself be accounted for in terms of natural ends. More generally, the appeal to natural ends as the touchstone for ethical theorizing confuses explanation with justification and in fact fails to generate answers to genuine moral perplexity. We can learn much that is relevant to ethics by attending to our biologically evolved nature, but through such attention, we can at most discover limits and disclose possibilities. The answers, if we can find them, must lie elsewhere.

CHAPTER TWO: DECISION THEORY AND INSTRUMENTAL REASONING

2.0 Introduction

Any constructivist project in ethics stands in need of some account of practical reasoning, of the way or ways in which reason bears upon action, and how action can be or fail to be rational.[1] Instrumental reasoning is a promising place to begin, and in decision theory, there is an impressive and impressively developed body of theory which is often thought to amount to a rigorous, formal and systematic treatment of instrumental rationality. If decision theory is really a systematization of ordinary instrumental reasoning, we can help ourselves to its methods and results; if not, we face further questions about the relation between the two and especially about which, if either, trumps the other in case of conflict.

[1] I shall speak interchangeably of what is rational and of what is reasonable, and correspondingly, of what is irrational and of what is unreasonable. I do not intend (as, e.g., Rawls does [1999, 315-317]) to mark any distinction between rationality and reasonableness.

What I shall claim is that decision theory does not map well onto ordinary instrumental reasoning, and moreover, that this is not due to defects in ordinary instrumental reasoning, but rather due to the failure of decision theory to capture some of its essential features. The power of decision theory for dealing with a variety of restricted contexts and specialized problems is considerable, but it does not amount to a general theory of instrumental reason. That is perhaps unsatisfying – one would like to have a rigorous and general theory – but I do not believe one has been worked out. It really is ordinary instrumental reasoning, disciplined but not displaced by theoretical results, with which we have to work.

I shall begin by circumscribing somewhat the field and issues I wish to address (I do not intend a comprehensive survey) and by settling some terminological points to facilitate discussion. Then, I shall outline the features of decision theory that will concern me and proceed to a more detailed examination. My principal concerns will be with maximization, the standing of the axiomatic conditions on preference, the force of decision-theoretic considerations for those whose preferences do not satisfy the axiomatic conditions, and the normative distinction between means and ends.

2.1 Preliminaries

Decision theory, so far as it will concern me, addresses itself to questions as to how it is rational to act, given a set of preferences, constraints and beliefs (or expectations).[2] About the preferences, we shall have more to say as we proceed. The

[2] Other applications, e.g., to modeling in economics and other social sciences, are beyond my scope.

Decision theory is sometimes contrasted with the closely related field of game theory by saying constraints are constituted by what the agent has to work with in acting or bringing his plans to fruition. The beliefs and expectations are those of the agent with respect to how the world is and what consequences will or are likely to follow upon different courses of action. These beliefs, except in the special cases in which, e.g., they are probabilistically incoherent, are generally taken as given by decision theorists and not, in any special way, within their competence. Whatever there is to be said about them with respect to whether the beliefs and expectations are correct or mistaken or as to whether they are arrived at in (generally) reliable ways, will involve other kinds of investigation (probably, many other kinds, since there is no unified field or discipline that addresses the correctness of what people take into account in making decisions).

Central to decision theory is the development and statement of conditions for the coherence of sets of preferences. No substantive conditions upon the content of preferences are assumed or presupposed, but sets of preferences, or sets of preferences in combination with sets of beliefs and expectations, can fail to be coherent. Conditions upon the coherence of preference sets are expressed by decision theorists in the form of axioms or postulates said to be definitive of the coherence of preference sets. Then, a rational choice is defined as one that is based upon – that is, either determined by or permitted by – a coherent preference set (again, in conjunction with beliefs and

that game theory is addressed to issues of strategic interaction, to how it is rational to act in interacting with other rational agents where (at least part of) what has to be taken into account is the fact that one is dealing with agents who can themselves take into account the fact that they are dealing with rational agents. The distinction can also be drawn in other ways, e.g., by describing game theory as the part of decision theory dealing with strategic interaction or viewing decision theory as the part of game theory that is confined to "games against nature," i.e., in which actions of and interactions with other rational agents are not relevant. The issues I wish to address do not require venturing into game theory, so I will set it aside. (They are still, however, relevant to game theory, since game theory embodies the same underlying conception of rationality.)

expectations, but I shall not keep repeating this).[3]

Different ways of axiomatizing the conditions on coherent preference have been offered. In some cases, these amount to different ways to reach the same destination. The different axiom systems turn out to impose equivalent conditions. In others, genuine alternatives to key axioms, such as Independence,[4] are proposed, and a preference set that qualifies as coherent under one axiomatization may not under another. As representative, I shall confine my discussion to standard expected utility theory.[5] There are two reasons. First, it is both the most familiar and the most widely relied upon; if anything deserves to be considered canonical in the field, standard expected utility theory is it. Second, the issues with which I am concerned emerge clearly in connection with it. To the extent that these issues are not raised by other axiomatizations, I see no need, as part of my current project, to address them.[6]

2.11 Matters of Preference

Some discussion of the key notion of preference is in order. What is a preference, and what kind of role does it play?

[3] A decision theorist might or might not allow that there is some other sense of rational choice that is not captured by whatever is the correct set of conditions on the coherence of preference sets. If he does, then he would treat the definition of rational choice in terms of the coherence of preference sets as defining some more restricted notion, rationality relative to preferences, perhaps.

[4] It is not important to define this here. I only note that Independence is critical for the possibility of defining a cardinal utility function (to be explained somewhat further later).

[5] Strictly, I shall be discussing both ordinal utility theory, which does not require the construction of an interval scale, and standard expected utility theory, which does. This is because ordinal utility theory is presupposed by expected utility theory (which is my major concern). For the present, I shall try to steer clear of the issues that arise when uncertainty (where information about probabilities is unknown or not available) as distinct from risk (where information about probabilities is available or assumed) affects the picture. Choice under uncertainty will occupy us at some length later.

[6] So far as the same issues *do* arise with other axiomatizations, what I say here will apply to those as well.

Broadly speaking, there are two ways preferences have been conceived by decision theorists. First – an approach which has been most popular among economists – preferences have been taken to be *revealed* in choice and action. Someone who selects one option over another is said to prefer the option selected. Choice of one option over another is taken to be *criterial* for the existence of a preference for the option selected over the other. Second, preferences have been conceived as mental states of some kind that underlie and explain choice behavior. Importantly, the second conception allows for some slippage between preference and choice-behavior: the fact that something is chosen is not sufficient (though it may be good evidence) to show that it is preferred.

I think the revealed preference interpretation is inadequate. One reason is that it fails to track what we ordinarily mean when we speak of preferences. For example, if preference is what is revealed in behavior, no one can ever be indifferent between or among a set of options.[7] But a sartorial unsophisticate such as myself may simply open the drawer and take the first shirt that comes to hand. I may prefer taking the first that comes to hand over some other selection procedure, but not the shirt I take to others I could equally well have taken. Between the shirts, I am indifferent.

Additionally, it seems that a person can have a preference that is never revealed in behavior because the occasion for it never arises. I may have a preference as to which car (beyond my present means) I would purchase if I won the Publishers' Clearing House Sweepstakes, but if I don't win, I will never get to select between them. If I don't actually select one, though, a consistent revealed preference theorist would have to say I

[7] Modifying a revealed preference account to say that choice reveals either that one prefers the option selected or is indifferent between them (equivalent to the notion of weak preference explained below) is not attractive, for then no evidence from choice behavior could show that one is not indifferent between all options.

have no preference between them.[8]

That ordinary usage is not captured by a revealed preference perspective is not decisive, however. There might be theoretical advantages to using the term in a specialized way. So, a more important and more fundamental objection is that adopting the revealed preference view makes it impossible to identify any sets of preferences as failing to meet the coherence conditions specified in decision-theoretic axiom systems. As David Friedman remarks (speaking of the use of assumptions about rationality in economics):

> In order to get very far with economics, one must assume not only that people have objectives but that their objectives are reasonably simple. Without that assumption, economics becomes an empty theory; any behavior, however peculiar, can be explained by assuming that the behavior itself was the objective. (Why did I stand on my head on the table while holding a burning $1,000 bill between my toes? I *wanted* to stand on my head on the table while holding a burning $1,000 bill between my toes.)[9]

The point is that any behavior can be represented as being in accord with preferences, provided that preferences are sufficiently weird, and so can any collection or sequence of behaviors. More generally, the coherence conditions of decision theory amount to

[8] The revealed preference theorist might instead hold that only choice behavior proves preference but deny that the absence of choice behavior proves the absence of preference. However, this can hardly be comfortable for him. Once preference is admitted to have some reality distinct from choice behavior and presumably at least partially available to introspective access, it will be hard to explain why choice behavior is always proof of a corresponding preference and why introspective reports to the contrary are always to be discounted.

[9] Friedman 1996, 4. Friedman is concerned with what assumptions about preferences and rationality are needed in order to make use of them for explanatory and predictive purposes. My concern is normative, but an analogous point applies.

imposing certain kinds of consistency requirements upon preference sets. The problem, if the requirements are to have any bite, is that no two (or three, etc.) behaviors that actually occur, considered apart from some description of them as, e.g., following a rule or aiming at an objective, can possibly be inconsistent with each other.[10] (If they were, they would not all occur.) If we adopt the revealed preference perspective, there are no obvious theoretical gains, and there is a significant theoretical loss: we lose the ability to distinguish between coherent and incoherent preference sets, and therefore, when rational choice is defined in terms of coherent preference sets, between rational and irrational choice.[11] So, I shall assume that preferences are mental states of some sort that underlie choice behavior and that we can (at least in some cases) know what a person's preferences are without awaiting their revelation in behavior.[12]

An important further point about preferences deserves our attention. When one speaks of ends or objectives (and similarly for goals, desires and wants), their content can typically be construed in terms of single-place predicates. There is some envisioned action or state of affairs that is the content of the objective; if the action occurs or the state of affairs comes about, the objective is realized. By contrast, preferences are

[10] For that matter, there is no inconsistency between doing *A* because it promotes *C* and doing *B* because it prevents *C* unless there is also some assumption to the effect that relevant preferences are the same at the times of the respective performances of *A* and *B*.

[11] I suspect this may have been an *attraction* for some. "Why," Nozick (1997, 133) asks, "does Mises [who provided an unusually explicit statement of a revealed-preference perspective, in his 1963, 19-21, 94-96, 102-104] think it is so important to argue that preferences cannot be irrational? Perhaps because he doesn't want anyone interfering with choices on the grounds that they arise from irrationally structured preferences." If that is the reason, the move has little to recommend it, for, in the first place, we might wonder why, if the bar is set so low that everything qualifies, it is either important or desirable to protect such "rational" choices from interference, and, in the second, any attempt to interfere would *also* qualify as rational.

[12] I do not intend to enter into any intricacies of the epistemology of preferences. I assume that agents have some (fallible) introspective access to the content of their preferences and that there are also various indirect ways of supporting or undermining claims about their preferences.

explicitly comparative and must be construed in terms of two-place predicates: *this* is preferred to *that*.

In many ways, this injects a salutary dose of realism. John Broome makes the point nicely:

> Imagine you meet a thirsty person. She wants water, she wants Coca-Cola, she wants beer, and she has a great many other wants too. Suppose you know all her wants in great detail. You know she wants half a pint of water; she wants a pint of water; she wants a litre of beer; she does not want Coke and beer together; and so on. Given all that, what should you give her to drink? A pint of water? A half-pint of Coke? Coke and water together? Just from knowing everything she wants, you cannot tell.
> To know what to give her, you need to know her *comparative* wants. You need to know what she wants more than what. You need to know her preferences, that is. Her preferences put all her options in an order: a pint of beer above a pint of water, a half-pint of Coke and half-pint of water above a half-pint of beer, and so on. If you know all her preferences, and if we grant [that she should be given what she prefers], you know what to give her; you should put her as high up her preference order as you can. But knowing just her wants is not enough. (1999a, 9-10)[13]

The real choices that people make are rarely matters of just realizing an objective

[13] "Similarly," Broome adds, "if you only know what is good for a person, even if you know everything that is good for her, that is useless. If you know everything that is good generally, that is useless too. You need to know what is better than what." (1999a, 10)

or not, with no other considerations brought to bear. Of course, we *do* take action to realize fairly definite objectives, but other considerations, almost inevitably comparative, are part of the background against which this or that objective is settled upon.

2.2 Outline of Utility Theory

With this background, I proceed to outlining utility theory.[14] So far as possible, I shall avoid technicalities, but it is useful to introduce some notation to capture the way in which preference is comparative. We can begin with the notion of *weak preference* ($\succeq$). One thing is weakly preferred to another when it is at least as good as (alternatively, no worse than) the other in terms of a preference ranking. "$A \succeq B$" should be read as "A is weakly preferred to B."

In terms of weak preference, we can define both *strict preference*[15] and *indifference*. A is strictly preferred ($\succ$) to B when it is definitely better in terms of a preference ranking – that is, when A is weakly preferred to B and B is not weakly preferred to A. Or, $A \succ B$ when $A \succeq B$ and it is not true that $B \succeq A$. A and B are indifferent ($\sim$) in terms of a preference ranking when it is true both that A is weakly preferred to B and that B is weakly preferred to A. Or, $A \sim B$ just in case $A \succeq B$ and $B \succeq A$.

The resemblance of this notation to that for comparing numerical or algebraic expressions ($>$, $\geq$, $=$, etc.) is, of course, not accidental. The idea of utility theory is to

[14] Largely, I am following Shaun Hargreaves Heap's treatment in his article, "Rationality," in Heap, *et al.* 1992, which in turn follows the approach of von Neumann and Morgenstern. In this approach, probabilities are assumed to be objective and given to the agent. In the alternative approach worked out by Savage and others (Savage 1973), probabilities are understood as subjective degrees of belief and attributed to the agent on the basis of choice behavior. For some useful discussion, see Found 2001.

[15] When I mean to be talking about a case in which one item is strictly preferred to another, I shall generally just say that it is preferred.

represent the relations between preferences by some kind of numerical index, so that certain important properties of the numerical relations will apply also to the relations between preferences.[16] Such an index can be used to represent a person's utility from the options available to her, and the property that it is important for such an index to have is that, from among her options (leaving aside ties), the largest number is assigned to the option she most prefers. Then, if she selects what she most prefers, since that is identical to selecting the option associated with the largest index number, she can be said to be maximizing utility (or, to anticipate coming elaborations, can be said to be maximizing expected utility).[17]

2.21 Ordinal Utility Theory

Plainly, for this to be workable, certain conditions on the relations between preferences will have to be satisfied. The numerical relations work because the numbers

[16] That there exists some orderly way of assigning index numbers to the elements of a set of preferences (which presupposes that the relations between the preferences themselves meet certain conditions, to be discussed further below) is what is meant by saying that a utility function can be specified (for that agent, with those preferences). In essence, a utility function amounts to a mapping of the elements of a preference set onto a number line, and the relations supposed to matter on the number line may be either cardinal or ordinal. Different utility functions – i.e., different mappings – may preserve the same set of relations between the elements of a preference set; thus, one correct mapping will be some transformation of other correct mappings.

[17] In order to avoid misleading suggestions, it is useful to remark briefly on the notion of utility employed here. The term 'utility' has historically been used in a variety of ways, and it is important not to confuse how it is employed in decision theory with others. What must be avoided is the idea that the utility maximized when the agent selects her best option is some separate object of her pursuit (perhaps a warm glow of satisfaction) distinct from other elements of her preference set, such as finding a job that matches her skills or a restaurant that serves good Thai cuisine. Naturally, it may be that some intra-psychic state, such as a warm glow of satisfaction, is among the elements of her preference set, but, if so, it will itself have to be assigned a utility index and will affect the utility indices that can be assigned to other elements. The warm glow will not be identical with her utility but will instead feed into assessments of the utility of her various options. To put it slightly differently, if the warm glow is an element in her preference set, it will have to be ranked *vis-à-vis* other elements as preferred, dispreferred or indifferent to them, and it may be that getting something else will have greater utility than the warm glow. Utility should not be understood as a separate object of pursuit but instead as a representation of preferences. To say that a person is maximizing utility is just to say that she is doing what she most prefers. See Broome 1999b.

they relate have certain properties. Similar requirements apply when transposed to relations between preferences.

For present purposes, the most important are Transitivity and Completeness. Assume an agent has a set of preferences over some set of elements. What Transitivity requires is that, for any three elements in the set, *A*, *B* and *C*, if *A* is preferred to *B* ($A \succ B$) and *B* is preferred to *C* ($B \succ C$), then *A* must be preferred to *C* ($A \succ C$).[18] If this condition were not satisfied, then she could regard *A* as (preferentially) better than *B*, but *C* as indifferent to or better than *A*.

What Completeness requires is that any element in the set can be ranked *vis-à-vis* any other. Take any two elements, *A* and *B*; then, it must be that the agent prefers *A* to *B* ($A \succ B$), that she prefers *B* to *A* ($B \succ A$) or that she is indifferent between them ($A \sim B$). If her rankings are not complete in this sense, then it would be possible that there is some pair of elements which are not comparable in terms of her preferences: neither is preferred to the other nor are they ranked equally. I think this is a real possibility, but for the present what is important is that if there is some pair of elements between which no preferential relation can be established, then, so far as the relation between those elements is relevant, it will not be well-defined what it is that best satisfies the agent's preferences (just as when, in numerical comparisons, we say that some term, such as *i*, the square root of negative one, has no place on the real number line, we cannot compare it to any real number and say, for example, that *i* is greater than, less than or equal to two).

When a set of preferences meets these conditions (and certain others[19]), then the

[18] The other relations, weak preference and indifference, must also sustain transitivity relations, but the strict preference relation is the most important.

[19] Strictly, two more conditions, Reflexivity and Continuity, are needed. Reflexivity requires that

elements over which the preferences range can be ordinally ranked – that is, they can be ranked from (preferentially) best to worst (including ties). We can then define an ordinal utility function for the preference set. This just means that we assign a number to each element in such a way that, for any two elements, *A* and *B*, (1) if $A \succ B$, then the number assigned to *A* is greater than the number assigned to *B*, and (2) if $A \sim B$, then the number assigned to *A* is equal to the number assigned to *B*. So, if Jennifer prefers beer to Coke, is indifferent between Coke and Pepsi, and prefers either Coke or Pepsi to water, we could assign six to beer, three to Coke, three to Pepsi and two to water. The particular numbers assigned do not matter, and in particular should not be taken to imply that Jennifer likes beer twice as well as Pepsi or Coke one-and-a-half times as well as water. Any other set of numbers that preserved the same ordinal relations – e.g., 903, 902, 902 and 107 – would serve equally well.

The number assigned to each element can be called its utility index or can be said to represent its (ordinal) utility. Assume now that the elements in an agent's preference set are outcomes of action that are known with certainty,[20] and that the preference set is complete and transitive. Then we can correlate each outcome with a corresponding action that is sufficient to bring it about. Assume also that these outcomes are all that

any element in a preference set be at least as good as itself. Continuity requires that, for any pair of goods in a bundle of goods, one of the two can be made marginally worse and the other marginally better in such a way that the resulting bundle would be judged to be, without change in preferences, indifferent to the first bundle. Satisfying Continuity would disallow any strictly lexical orderings of preferences. That is, in a preference set satisfying Continuity, it cannot be the case that there is no quantity of a good, *B*, that would compensate for some small reduction in another good, *A*. Continuity does not, however, rule out preference-orderings that are practically equivalent to lexical orderings. (There are actually two different Continuity axioms, the one just explained, which is needed for ordinal utility theory, and the other for its extension to expected utility theory. I briefly discuss the other Continuity axiom later.)

[20] Of course, a person may have preferences over elements that are not outcomes of action (at least for that person) at all, such as whether a scientific theory is true or whether a past event happened in a certain way. If such preferences have no action-guiding import – if they do not, for example, shape the course of an investigation – then they can be set aside as not bearing on the rationality of action.

matter – in particular, that no preferences with respect to the actions as distinct from the outcomes affect what the agent would, all things considered, prefer.[21] If these conditions hold, we can give content to the earlier claim that a rational choice is based upon a coherent preference ordering and can say unambiguously what action is best in terms of the agent's preferences. We assign numbers to each outcome in a way that preserves the right relationships, and then the action correlated with the outcome with the largest utility index (or one of them, in case of ties) is the one that would best satisfy her preferences. Her best choice is the one that maximizes utility – or, in other words, is one that selects an outcome with the largest associated utility index.

2.22 Expected Utility Theory

Though ordinal utility theory may be useful for dealing with preferences over certain outcomes, it has significant limitations in that, in most of our choices and deliberations, we do not know with certainty what the outcomes of our actions will be. Thus, even if an agent has a complete and transitive preference ordering over outcomes, that is not generally sufficient to make it clear what action will best satisfy his preferences (or what action to take given that he cannot be sure the actual outcome will best satisfy his preferences).

What needs to be done to extend utility theory so that it has application in a world

[21] This can often be secured by 'loading' preferences with respect to actions into the descriptions of the associated outcomes. For example, it may mislead to say that I am comparing a mowed to an unmowed lawn. The right description might be that I am comparing having a mowed lawn, together with being hot and sweaty, to having an unmowed lawn, together with being cool and comfortable.

Such loading of preferences with respect to actions into outcome-descriptions is not always possible nor is it always clear, where it is not possible, that the agent whose preferences resist such redescription is being less than rational – see Hampton 1998, Chapter 8 and Verbeek 1999 – but I shall, for the present, set aside such problems.

of risk? I will begin with an assumption about risk and then follow up with two terminological points. The assumption is that the risk an agent faces in selecting from among his options can be characterized in terms of probabilities which he knows or takes as given. If there is some outcome which the agent would prefer to all others, then he knows for each of his options that it has some definite probability, a point-probability as it is called, of leading to the outcome he prefers and also has some definite probability or probabilities of leading to some member of a set of relatively dispreferred alternative outcomes.[22]

With regard to terminology, I have spoken frequently of the *elements* of a preference set as what preferences range over. The term was selected deliberately to avoid any pre-judgment as to what preferences could range over. In ordinal utility theory, it is most convenient to assume that the relevant preferences range over certain outcomes, but there is no necessity for this.[23] An initial step to incorporating risk would be to replace talk of certain outcomes with talk of *prospects*, which may be defined as outcomes combined with a measure of their probability.[24] More precisely, any prospect, *A*, is a gamble in which some outcome, *B*, is received with some probability, *p*, and the alternative to *B*, not-*B*, is received with the complementary probability, 1 - *p*. A complication to be borne in mind is that the "outcomes," *B* and not-*B*, may themselves be

[22] This does not seem to be the only way in which an agent can be ignorant of what the future holds (see note 5), but one problem at a time!

[23] Indeed, in an earlier note (20), I mentioned the possibility that one might have preferences about which theories are true, etc.

[24] To avoid cumbersome locutions, I shall generally speak of *certain* and *risky* prospects, where a certain prospect is one in which an outcome is assigned a one-hundred-percent probability (this could also be called a non-compound prospect), and a risky prospect is one in which an outcome is paired with some probability less than one hundred percent. This terminology – as contrasted with the more natural 'uncertain prospects' – is adopted with a view to avoiding confusion over the distinction mentioned earlier (note 5) between uncertainty and risk.

prospects, embodying further gambles over outcomes and so on. So construed, outcomes that are certain are a subset of prospects, namely those in which some outcome, that does not itself involve any further gamble, has a probability of unity.

Second, we can also generalize the notion of utility and replace it with *expected utility*, which can be understood as a representation of an agent's preferences over prospects. We can stipulate that, whatever the utility of an outcome received with certainty is, that is also the *expected* utility of the corresponding prospect – that is, one in which the relevant outcome has a hundred-percent probability of occurrence. Thus, in parallel to the above treatment of the relation between prospects and outcomes, the utilities of certain prospects are special cases of the expected utilities of prospects. This, however, does not get us very far. In particular, it does not license any inferences that the prospect in which an outcome has a fifty-percent probability has an expected utility equal to half that assigned to the prospect in which the same outcome has a hundred-percent probability. The reason is that the selection of a particular number to represent the utility of a given outcome in ordinal utility theory is quite arbitrary, so long as certain relations are preserved with the index numbers assigned to other outcomes. Thus, two equally good ordinal utility functions representing the same set of coherent preferences might exhibit very different proportional relations among its elements. If we cannot treat Jennifer getting a beer as satisfying her preferences twice as well as her getting a Coke, we also cannot treat her having a fifty percent chance of getting a beer as being just as good as getting a Coke. (That is, we cannot without further information about her preferences and their interrelations.)

If we are to deal with this within the framework of utility theory, some way needs to be found of further regimenting the numerical representation of preferences and their relations to one another beyond the ordinal relations already allowed for. Somehow, we must find a way of cardinally scaling the numerical relations between representations of preferences. To get a better grasp on what is needed, let us look a bit more closely at Jennifer's problem. Assume that only her preferences with respect to what to drink are relevant and that getting a Coke, a beer or remaining thirsty are the only options. She would rather have a beer than a Coke and would rather have either than remain as she is (namely, thirsty), so getting a beer for certain has to be assigned a greater expected utility than getting a Coke for certain, and the certainty of getting a Coke will itself be assigned a greater expected utility than will be assigned to the certainty of remaining thirsty. A risky prospect of getting either a beer or else remaining thirsty will have some intermediate value, but it is not yet clear how that intermediate value will relate either to the certainty of getting a Coke or to any of the infinitely many possible risky prospects in which she either gets a Coke or remains thirsty.[25]

Now, it is initially plausible that answers to questions of the type just suggested can be reached on the basis of the agent's preferences, that preferences have *degrees of strength* with respect to one another rather than just being ordinal. Jennifer may say that she would *very much* rather have a beer than a Coke. It is also plausible that these degrees to which one thing is preferred to another extend to the ordering of risky prospects. If Jennifer much prefers having a beer to a Coke, then it is likely that she would also prefer, say, a ninety-nine percent chance of getting a beer to the certainty of

[25] I am simplifying by considering only cases in which improvements over the *status quo* (without worsenings) are being considered. Thus, a risky prospect of an improvement over the *status quo* should also count as an improvement. If worsenings from the *status quo* were under consideration, then a risky prospect of a worsening (without any prospect of an improvement) should also count as a worsening, but a lesser one than the certain prospect of a worsening.

getting a Coke. What is needed, then, is to establish a cardinal scale along which preferences can be represented that (a) preserves the ordinal relations between the certain or non-compound prospects in the preference set, and (b) also allows the comparison and ranking of any prospect, whether certain or not, with respect to any other.[26]

It turns out, if certain further conditions are imposed upon a preference set, that this can be done. The intuitive idea can be presented fairly readily. We arbitrarily assign numbers to a worst and to a best outcome for an agent (with the larger number being assigned to the best outcome, of course), which are, respectively, worse than or better than any of the actual prospects we wish to compare. Call the best *Bliss* and the worst, *Torture*. It should be understood that whatever Torture is, it is so bad that the agent, if given a choice between Torture and anything else, would select whatever is the alternative to Torture. Similarly, Bliss is so good that, when compared to anything else, it would be selected over that alternative. If we assign, say, the number zero to Torture and one hundred to Bliss, it is plausible that all the agent's other preferences, whether for certain or risky prospects, can be arrayed along a number line, preserving their ordinal relations to one another, somewhere between zero and one hundred. But where should each one – Jennifer's preference for getting a Coke, for example – be placed? Well, since Coke is equivalent neither to Bliss nor to Torture, we have (using obvious abbreviations) the following relation:

$$B \succ C \succ T$$

Since numbers have been assigned to Bliss and Torture, can we assign a number

[26] Cardinal measures and comparisons do not presuppose a non-arbitrary zero-point (such as zero mass for cardinal comparisons of mass). They can be constructed for domains in which there either is not or is not assumed to be such a zero-point (as temperature scales were constructed before it was recognized that there is an absolute zero).

to Jennifer's getting a Coke as some function of her preferences between Bliss and Torture? The first step is to recognize that there does seem to be a way of combining her preferences with respect to Bliss and Torture so as to get an intermediate value. Specifically, we can construct a risky prospect, a gamble, consisting of a probability mix of the two, where she gets Bliss with some non-zero probability and Torture with the complementary probability. That is, we say that the gamble gives her Bliss with probability, p, and Torture with probability, $1 - p$ (where $p \neq 0$). We can symbolize this gamble as:

$$[B, p; T, 1 - p]$$

This gamble has to be preferred to Torture and dispreferred to Bliss, because having some chance of Bliss has got to be better than the certainty of Torture and some chance of Torture must be worse than the certainty of Bliss. This is put to use in two ways. The first is to calibrate the scale between Torture and Bliss. Assume that Jennifer is considering two different probability mixes between Bliss and Torture, $[B, p; T, 1 - p]$ and $[B, p^*; T, 1 - p^*]$, where $p^* > p$ – that is, in which the second gamble gives her a greater probability of getting Bliss and a smaller probability of getting Torture than the first. It seems reasonable to suppose that she will prefer the second gamble to the first:

$$[B, p^*; T, 1 - p^*] \succ [B, p; T, 1 - p]$$

If this holds true for all values of p and p^*[27], then, for any value of p in $[B, p; T, 1 - p]$, we can assign to that gamble the real number along the zero-to-one-hundred scale that is equal to p x 100.[28] Then, every point along the scale will correspond to a value for p that

[27] Again, where $p^* > p$.

[28] It was not accidental that I selected a zero-to-one-hundred scale. Those numbers were arbitrary in that others could have been used to anchor the end-points, but using zero for Torture and one hundred for

itself appears in one and only one particular gamble of the form, $[B, p; T, 1 - p]$. Moreover, if we use these points along the scale to provide expected utility index numbers for the corresponding gambles, then all of the infinitely many possible gambles between Torture and Bliss will stand in the right ordinal relations to one another.[29] Every gamble that gives Jennifer a greater probability of Bliss (and a smaller probability of Torture) than some other will be assigned a larger index number than that other.

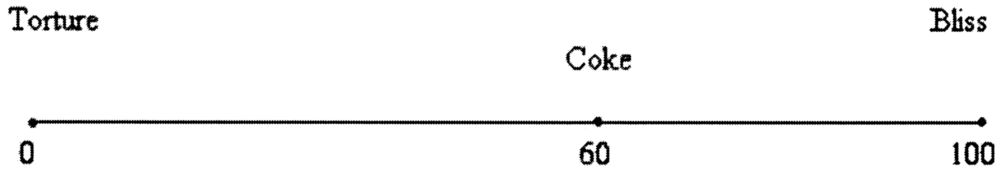
Once we have calibrated the scale in this way, the second use is to find out where along it to represent a preference for some particular prospect other than a gamble between the end-points, such as Jennifer's preference for getting a Coke. At this point, the idea is that, since we know a given probability mix between Torture and Bliss is, just like getting a Coke, preferred to Torture and dispreferred to Bliss, we can consult Jennifer and ask her whether she would prefer to have this gamble, for some specific value of p, or the certain prospect of getting a Coke. If she would rather have the Coke than accept the gamble, then the p-value is too small to represent her preference for getting a Coke. If she would prefer the gamble to the Coke, then the p-value is too large. In principle, if we present her with a large number of gambles between Torture and Bliss, we should be able to find one with respect to which she is indifferent between accepting that gamble and getting the Coke.[30] Then, the number on the Torture-Bliss scale that corresponds to that

Bliss simplifies the exposition.

[29] This is not the only possible way to secure these properties, just one that is simple and intuitive.

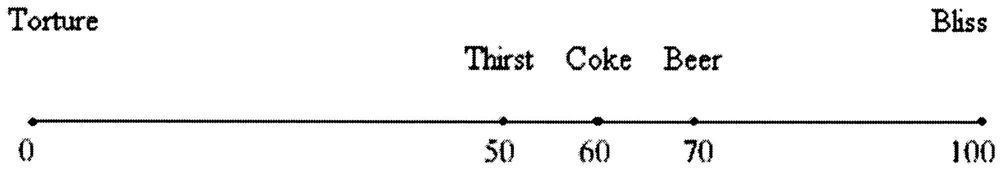
[30] The "in principle" clause here is important since it may well be that Jennifer is getting more and more thirsty as we ask her questions about gambles between Torture and Bliss with differing p-values (rather than giving her a Coke!). As her thirst approximates Torture, she may be willing to accept gambles with lower p-values as being indifferent to getting a Coke. To be precise, we would have to say instead that, in a given situation, with given levels of thirst, etc., there is some gamble between Torture and Bliss, such that, *if* she were offered it, she would be indifferent between accepting it and the certain prospect of a Coke.

gamble can be assigned to represent her expected utility in getting a Coke. Then, we may get a graphic representation of the result that looks like this:



What this would mean is that, for Jennifer, C ~ [*B*, .6; *T*, .4] – that is, that she is indifferent between getting a Coke for sure and a gamble in which she has a 60 percent chance of Bliss and a 40 percent chance of Torture.

We repeat the process for Jennifer's other preferences, such as those for getting a beer or remaining thirsty, determining at just what gambles between Torture and Bliss she would be indifferent between those gambles and the respective certain prospects in which she remains thirsty or gets a beer, to find expected utility indices for those prospects as well. Suppose that, when we have done so, the slightly enriched graphic representation looks like this:



If we assume that Jennifer is thirsty and if all has gone well, we can not only rank her thirst, her getting a Coke and her getting a beer ordinally against one another, we can also determine, for example, that she would be indifferent between getting a Coke and a 50 percent chance of getting a beer or that she would prefer a 60 percent chance of getting a beer to the certainty of getting a Coke.

Generalizing a bit (and assuming the scale is worked out in the necessary detail), we can represent on the scale every certain prospect she faces. Further, we can represent all her preferences with respect to risky prospects as functions of the corresponding certain prospects on the same scale, and we can rank each prospect, whether certain or not, with respect to every other and assign to each an expected utility index. The index numbers assigned will have the property that any prospect strictly preferred to another has a larger expected utility index than the one to which it is preferred. Thus, Jennifer's choices, if she is actually selecting what she most prefers, maximize her expected utility. Further, by virtue of the possibility of constructing an expected utility scale that allows ratio comparisons, that is, that allows us to say to what comparative *degree* her preferences are satisfied by different prospects, we are able to extend the reach of the idea that a rational choice is one that is determined or permitted by a coherent preference set to cover cases in which outcomes are not known with certainty.

Now, there are at least two reasons for being suspicious of what I have been doing in the last several paragraphs. The first is that I may have drastically over-simplified the kinds of options over which preferences range. It is not realistic to speak, say, of Jennifer's preferences with respect to beer as though it makes no difference what quality, temperature and quantity of beer is on offer, and so on. Nor is it enough to further specify in isolation the characteristics of the beer, for her preferences with respect to beer are also affected by how thirsty she is and what else she cares about in the present circumstances.

This problem, however, is, in principle, easily remedied. What has to be realized is just that her preferences for beer, Coke and so on need a more fine-grained description if we are to be sure that *these* are the preferences which apply to her current situation. She need not have *general* preferences for beer over Coke over thirst. It suffices if she has preferences for beer (with the available characteristics) over Coke (with the available characteristics) over her actual degree of thirst in the current circumstances, when these highly specific preferences can be assigned expected utility indices.

The second concern is much more serious. I have alluded to further conditions on the coherence of preference sets, but have actually said very little about these further conditions. Instead, I have engaged in a good bit of hand-waving in the form of assertions about what is plausible, what it is reasonable to suppose and the like. But *is* what I claimed plausible or reasonable to suppose? In particular, is it plausible or reasonable to suppose that they embody or exemplify *requirements* for the coherence of preference sets and therefore, since decision theorists define rational choice in terms of the coherence of the preference set from which it proceeds,[31] *requirements* for rational choice? To consider this question is a way of raising the issue whether and why the axiomatic conditions upon the coherence of preference sets are normative for choice. And that can hardly be approached without saying what the further conditions are.

A natural way to proceed at this point would be to spell out the further axiomatic conditions and then undertake to assess them, considering arguments in their favor as well as alleged counter-examples, but that is a well-trodden path and I shall take a different tack. I shall indeed *briefly* say something further, with only a minimum of

[31] See note 3 and accompanying text.

commentary, about what the axiomatic conditions are.[32] But then I shall consider their normative standing from a different angle.

2.221 Additional Axiomatic Conditions for Expected Utility Theory

Four conditions have already been introduced in the discussion of ordinal utility theory: Reflexivity, Transitivity, Completeness and Continuity.[33] Six additional conditions can be identified.[34]

The first of these can be called *Extension to Prospects*. What it says is that the first four conditions – Reflexivity, Transitivity, Completeness and Continuity – must also be satisfied by sets of preferences ranging over prospects. This seems just as reasonable as the initial conditions themselves since, when those were introduced, they were specified as applying to the ordering of elements of sets of preferences. To say that they apply to prospects is just to recognize prospects as elements of preference sets.

The second condition can be called *Preference Increasing with Probability*. What it requires is that for any pair of prospects, *A* and *B*, if *A* is preferred to *B*, then, for any pair of gambles over *A* and *B*, the gamble in which *A* is assigned a higher probability (and *B* a lower probability) is to be preferred to the other. That is, if $A \succ B$, then $[A, p; B, 1 - p] \succ [A, p^*; B, 1 - p^*]$ if and only if $p > p^*$. The idea is just that if some prospect is

[32] I shall try to keep technicality also to a minimum, but some is almost unavoidable.

[33] Transitivity and Continuity, as introduced, were defined in terms of strict preference relations, but analogous conditions on weak preference and indifference must also hold. Similar analogous conditions are needed for the remaining conditions. (Technically, it is more elegant to state the conditions in terms of weak preference and derive the requirements for strict preference and indifference as needed. However, stating the conditions in terms of strict preference is more intuitive, since "strict preference" is just what is ordinarily meant by "preference.")

[34] This is somewhat arbitrary. As noted earlier, there are different ways of axiomatizing decision theory, and what exactly the conditions are, and therefore what the precise number of necessary conditions is, need not be the same in different axiomatizations.

preferred to an alternative, then a larger chance of getting the prospect one prefers, rather than its alternative, has to be preferred to a smaller chance.

The third condition can be called *Closure*. What Closure requires is that, for any two prospects that are elements of a preference set, A and B, then, for any value of p, the gamble between them, $[A, p; B, 1 - p]$, is also an element of that preference set. That is, the elements of a preference set include any prospects that can be constructed from a gamble over other elements of the preference set.[35]

The fourth condition is sometimes called Continuity – somewhat confusingly, since there is already a Continuity axiom. I shall call it *Probabilistic Continuity*. Probabilistic Continuity says that for any three prospects, A, B and C, if A is preferred to B and B to C, then there is some probability-value, p, such that the gamble between A and C, $[A, p; C, 1 - p]$, is indifferent to B.

The fifth condition can be called *Strong Independence*.[36] It "means that, in any prospect, any component object or prospect can be replaced by an object or prospect indifferent to it, and there will be indifference between the resulting prospects and the original one."[37] In other words, what it requires is that an agent who is faced with a gamble between two prospects, A and B, and for whom B is indifferent to some third prospect, C, should be indifferent between the gamble, $[A, p; B, 1 - p]$, and the gamble, $[A, p; C, 1 - p]$. Since prospects involve gambles over outcomes that may themselves involve gambles over further outcomes, this means that more or less complicated

[35] Closure is included for expository convenience, but is not strictly needed since it can be derived from the other axioms.

[36] There are other, weaker, Independence axioms that figure in different axiomatizations. Where weaker versions of Independence are used, other axioms have to be strengthened to compensate.

[37] Heap, *et al.* 1992, 10.

gambles, between which the agent is indifferent, may be interchanged for one another in a course of deliberation and that refusal to accept such interchange of indifferent prospects marks a preference set as incoherent.

Finally, it is required that probabilities be combined in the normal way. Suppose that there are three prospects, A, B and C, and a compound gamble between them, $[[A, p; B, 1 - p], p^* ; C, 1- p^*]$. Then the probability of getting A is equal to $p \times p^*$, the probability of getting B is equal to $(1 - p) \times p^*$, and the probability of getting C is equal to $1 - p^*$. Accordingly, if there is some expected utility index assigned to each of A, B and C – represented respectively as $u(A)$, $u(B)$ and $u(C)$ – then the expected utility of the gamble is equal to $((u(A) \times p) \times p^*) + ((u(B) \times (1 - p)) \times p^*) + (u(C) \times (1 - p^*))$.

2.3 Decision Theory: Some Limitations

Much can be said about the various axiomatic conditions, taken one by one. Some, such as Transitivity, are almost universally accepted, while others, such as the Independence requirement, have attracted considerable suspicion.[38] What I shall try to do, however, is to press a somewhat different question, why the axiomatic conditions on preference sets should be taken to be normative for choice.

An entry point for considering that question is to remember that the source of the mathematics used to define cardinal utility functions is in measurement theory. It had been shown that, in any domain of elements with certain properties, properties which can be specified by a set of axioms, cardinal measures and therefore cardinal comparisons could be defined. What von Neumann and Morgenstern did, in working out the

[38] Still, it remains true that some form of Independence is accepted by most decision theorists.

mathematics of expected utility theory, was to transpose that set of conditions to the domain of preference. Thus, they were able to show how a set of preferences could be understood to meet the axiomatic conditions and so, how it was possible to develop a cardinal measure in terms of which preferences could be ranked and compared with one another.[39]

Strictly, what had been shown was that meeting the conditions is sufficient for it to be possible to define a cardinal measure. So far as I know, the additional claim that meeting those conditions is also necessary for a cardinal measure has not been proven at all, but, for the sake of argument, let us suppose that to be true as well.[40] If it is, then of course it has to apply to sets of preferences, so the axiomatic conditions on preference sets are necessary to define a cardinal expected utility measure.[41] But why exactly are the conditions necessary to establish a cardinal measure also normative for choice? Or, to put matters the other way around, why suppose that action in accordance with a preference set for which no cardinal measure can be defined is *rationally* defective?

So far, this is only a question intended to raise a doubt, but it can be fleshed out a bit. Consider the following argument[42]:

[39] Dawes 1988, 150-151.

[40] If it is *not* true, then the case for saying that the (assumed) axiomatic conditions on preference sets are normative for choice is weakened, for there might be some other set of conditions that the same preference set satisfies that is also sufficient to define a cardinal measure. If we call the standard set of conditions, A_1, and an alternative, A_2, then it might be that some action (outcome, etc.) best satisfies preference in terms of A_1 but not in terms of A_2, or *vice versa*. It is also possible that, though satisfaction of the conditions of either A_1 or A_2 is sufficient to cardinally order a preference set, a given preference set may satisfy one without satisfying the other. Then, the question would arise as to which, if either, is normative for choice.

[41] Though not proven, this is made plausible by the fact that no one seems to have any idea how one might go about defining a cardinal measure when one or more of the axiomatic conditions is unsatisfied.

[42] This argument assumes that rational choice must range over cases in which risky prospects are relevant. Since some measure of ignorance as to what the future holds almost always infects the options

(1) Unless the axiomatic conditions on preference sets are satisfied, we cannot have a cardinal measure of preference satisfaction.

(2) Unless there is a cardinal measure of preference satisfaction, there cannot be a determinate answer as to what best satisfies preference.[43]

(3) Therefore, there can be no rational choice (with respect to a preference set[44]) unless the axiomatic conditions are satisfied.

(4) And therefore, if rational choice is possible, the axiomatic conditions must be satisfied.

Plainly, those conclusions do not follow from the first two premises alone (which, for the present, I am granting). The first two premises could be true and the conclusions false if there could be a rational choice that does not determinately best satisfy preferences. The additional premise needed to derive the conclusion, (3), and the equivalent (4), would be something like:

among which we choose, this amounts to little more than saying that rational choice must apply to the options we face. (We *could* have a conception of rational choice which applies only to few or none of the choices with which we are actually faced, but it would be hard to explain why we should be much interested in it.)

[43] I stipulate that "what best satisfies preference" means "what satisfies preference at least as well as any alternative." In other words, ties are permitted at the top of a preference ordering. In the case of a tie for what best satisfies preference, any of the tied items may be chosen and would count as best satisfying preference. Additionally, for there to be a *determinate* answer as to which of a set of alternatives best satisfies preference, it must be true that an alternative that best satisfies preference is at least weakly preferred to any other. (There could be indeterminacy at the top of a preference ranking if there were a set of two or more options which could not be ranked with respect to the other[s], but each member of the set was ranked more highly than any option that was not a member of the set. This qualification is added to foreclose the interpretation that members of such a set of mutually unranked options could be said to satisfy preference at least as well as any alternative.)

[44] I shall not continue to repeat this qualification.

> 2a. Unless there is a determinate answer as to what best satisfies preference, there can be no rational choice.

That premise, however, admits at least three interpretations. The weakest is:

> 2a'. Unless there is a determinate answer as to which of a set of options best satisfies preference, there can be no rational choice among those options.

So understood, the premise may be unexceptionable. It can be taken to assert just that if the members of a set of options cannot be ranked as preferentially better than, worse than or equal to one another, there can be no decision between them on the basis of preference. But this is too weak to support either (3) or (4) in two respects. First, it does not imply that there is anything *irrational* about selecting such an option.[45] Second, it does not imply that members of the set cannot be ranked in terms of preferences against other options that are not a part of the set. There might be some option that is definitely better than any member of the set; if so, then it would be irrational to select a member of the set rather than that option. Or there might be some option definitely worse than any member of the set, in which case it would be irrational to select it rather than a member of the set. There could still be rationally required choice, even if not *among* the members of the set.

[45] It is relevant here that there are two different senses of 'rational.' It may, on one hand, mean *rationally required*. In that sense, there can be no preference-based rational choice among options that are not preferentially ranked. (For that matter, there can be no rational choice in *that* sense among options that are preferentially ranked, but ranked as indifferent to one another.) Or, 'rational' may mean *rationally permitted*. If rational choice is rationally permitted choice, then the selection of one from among a set of options that are not preferentially ranked may be rational. A case that selection of one of a set of preferentially unranked options is irrational would involve the claim that so choosing would be *rationally forbidden* or, alternatively, not rationally permitted.

In other words, (2a') is consistent with the existence of a partial ordering of the elements of a preference set.

So, consider a stronger interpretation:

> 2a". Unless there is a determinate answer as to which of a set of options best satisfies preference, there can be no rational choice among any options.

This is sufficiently strong to support (3) and (4), but, as it stands, it is implausible. For suppose there is a pair of options, *B* and *C*, which cannot be ranked preferentially against one another. From (2a'), and therefore from the stronger (2a"), it follows that there can be no preference-based reason for selecting one over the other. But suppose that the actual decision problem facing an agent is between *A* and *D*, that *A* is strictly preferred to *D*, and that *B* and *C* do not figure in his deliberations at all. Surely, then he could, and rationally should, select *A* over *D*. The fact that there are other options which he is unable to rank with respect to each other – options that, as it happens, he does not have to choose between – should make no difference. Once again, the possibility of a partial ordering of the elements of a preference set undermines the conclusions.

What is needed to support (3) and (4), then, is not precisely a stronger interpretation – (2a") is already sufficiently strong – but something that makes explicit a claim that (2a") only presupposes:

> 2a‴. Unless there is always a determinate answer as to which of any set of options best satisfies preference, there is never a determinate answer, and so, there can be no rational choice among any options.

What (2a‴) rules out is the possibility that a set of preferences may be partially ordered. In effect, it asserts that a partial ordering is no ordering at all. I want to draw attention to two closely connected features, one almost explicit and the other implicit in (2a‴). The first is that rational choice depends upon there being an ordering of options, combined with the further claim that only a complete ordering is genuinely an ordering. Obviously, this is closely connected with the Completeness axiom.

But why should this be accepted? This leads to the second feature. The implicit answer to this question depends on the assumption that rational choice requires or is to be identified with *maximizing*, with selecting what is best in terms of all of one's preferences taken together. And the possibility of doing that depends upon one's preferences being such that every element in one's preference set can be unambiguously ranked against every other.[46] When the elements of a preference set include prospects, that is only possible for practical purposes if a cardinal measure of preference satisfaction can be defined. Perhaps an infinite mind could (just) ordinally rank all prospects, including those defined as gambles over other prospects.[47] But no finite mind could follow suit. Suppose some agent prefers *A* to *B*, *B* to *C*, and *C* to *D*. How, without a cardinal

[46] Unambiguous ranking carries with it, of course, the requirement of Transitivity. If the ranking is not transitive, then, for at least some cases, the same option will be ranked as both better and worse than some other, depending upon the order in which it is compared with other options.

[47] Strictly, a genuinely complete and transitive ordinal ranking of all prospects could always be represented by some cardinal function or other, but it would not need to be the case that any particular ranking of some subset of prospects was derived from the cardinal function.

measure, is she to rank the gambles, $[A, p; C, 1 - p]$ and $[B, p^*; D, 1 - p^*]$, where $0 < p < 1$ and $0 < p^* < 1$? A preference for *A* over *B* over *C* over *D* would not imply any particular preferential relation between the specified pair of gambles. The only remotely tractable procedure for a finite mind is to treat preferences over gambles as a function of preferences over outcomes comparable on a cardinal scale.

The earlier argument, then, can be reformulated in this way:

(1) Rational choice presupposes maximizing.

(2) Maximizing presupposes the possibility of a cardinal measure of preference satisfaction.

(3) The possibility of a cardinal measure of preference satisfaction presupposes that the axiomatic conditions upon the coherence of preference sets are satisfied.

(4) Therefore, rational choice presupposes that the axiomatic conditions upon the coherence of preference sets are satisfied.

That argument is certainly valid, but since I am willing to grant the second and third premises, everything important in it depends upon the truth of the first. What I shall do in much of the remainder of the chapter is to focus in various ways upon its truth and upon that of the closely related requirement that one's preferences completely and unambiguously order all of one's options. My general strategy will be to run the argument in reverse: to argue that it is implausible that our preferences completely order our options and therefore implausible that rationality requires *of us* that our actions be determined or permitted by a preference set that satisfies the axiomatic conditions. That strategy would have no prospect of success, however, if there were some other argument showing that satisfaction of the axiomatic conditions on preference sets was necessary for rational choice, so the first matter to attend to is whether any such necessity *is* established by other arguments.

2.31 Defending the Axiomatic Conditions: Three Approaches

Nobody supposes that the axiomatic conditions on preference sets amount to logically necessary truths about rational choice or, therefore, that it is contradictory to deny or reject one or more of them. Nor is it supposed that it would be contradictory to deny one while keeping the others. If it were, then the one denied could be derived from the others and so would not be needed as an independent axiom.[48] The actual defenses offered for the axioms fall into three classes, which are not often clearly distinguished from one another.[49] First, it may be claimed that the axioms are, individually, intuitively secure or compelling or that they can be derived from something which is. Second, a coherentist defense can be mounted to the effect that accepting the set of axiomatic conditions makes the best sense of our pre-theoretic practice and of our assumptions and convictions about rational choice. Third, and most interestingly, a pragmatic case can be

[48] This may be slightly misleading. There are first, as has been noted, different ways of axiomatizing expected utility theory and some may be more or less compact in terms of the number of axioms relied upon than others. (For example, I included Closure among the axioms, though it can be derived from the others.) Second, there is theoretical interest in seeing what the most compact or minimal set of axioms necessary is. However, it remains true that, for whatever is the most compact set of axioms, each of them (and each proper subset of the axioms) is logically independent of and therefore not derivable from the rest.

[49] I am idealizing in describing pure cases of each of the approaches. It may be doubted whether anyone advocates taking any one of these approaches, to the exclusion of the others, to defending all of the axioms. I shall later briefly address the possibility that the best defense of the axioms consists of some appropriate combination of the different approaches.

developed that we will do better if our preferences and choices conform to the axioms.

2.311 The First Approach: Intuitive Security

A defense in terms of the intuitive security of the axioms runs into two sorts of empirical problems. One, which has been demonstrated by various psychological studies, is that situations can be constructed in which people systematically make choices that violate the axiomatic conditions. By now, there is a large literature on such violations.[50] Regularly and predictably, people choose in ways in which they would not if their preferences conformed to the axioms. It is an important fact that these violations are systematic. It is not just that mistakes are made about what is rationally preferable. That might be explained in any of several ways, including lack of time to work out what is best, insufficient familiarity with relevant procedures and the like. Then, however, one would expect a random distribution around the correct answer. But when the violations of the expected utility axioms are systematic – when test subjects tend to converge upon the *same* alternative to what would be required by the axioms – that suggests that something deeper, such as the common rejection of some axiomatic condition, is responsible for test-subjects' choices. Moreover, the systematic violations persist to some substantial degree even when the decision-theoretic arguments for an alternative choice are explained. The fact that the violations are systematic, even on the part of test-subjects who have been exposed to the arguments for an alternative, suggests that most people do not find the standard axiomatic conditions to be intuitively secure or

[50] See, e.g., Camerer 1995 and Kahneman and Tversky 1990.

compelling nor do they see them as flowing from something which is.[51]

The second empirical challenge is a more specialized version of the first. It can be called the challenge of the experts. Significant numbers of people who have as good a claim as anyone to expertise in decision theory find that, in certain decision problems, they are inclined to choose in ways that violate some axiom, usually some version of Independence. This inclination is often reflectively stable in people who have considered and understood all that can be said on behalf of conforming to the axioms.[52] Even if it might be said with some color of plausibility to be unsurprising if ordinary people, untrained in decision theory, have unreliable intuitions, it is far more difficult to make it credible that experts are unreliable in the same way. An appeal to the intuitive security of the axioms seems to lead only to a contest of divergent intuitions.

2.312 The Second Approach: Coherentist Defenses

The second or coherentist defense seems to be in no better shape, and for essentially the same reasons. As E.F. McClennen puts it:

> What one hopes for ... are starting points that command nearly unanimous acceptance, at least among thoughtful and knowledgeable researchers. Unfortunately, [the axiomatic conditions] do not appear to meet this test. [They] have been the subject of sustained, spirited and thoughtful

[51] An analogy: Suppose a simple arithmetical problem, such as 24 + 27, were posed to an elementary school class. It would not be terribly surprising, and would be less surprising the younger the students, if a large percentage got the answer wrong. But it would demand explanation in terms other than simple error if the majority of the class agreed upon a *particular* mistaken answer, such as 45. The need for such explanation would be still more patent if it was explained to the class why the answer was 51 and most of them *still* gave 45 as the answer.

[52] Two important examples can be found in Allais 1990/1979 and Ellsberg 1990/1961.

> questioning by a number of decision theorists. Thus, they appear to be unsuitable starting points. But this last consideration would also seem to work against any "coherentist" argument as well. The principles in question do not codify the choice behavior of competent or even expert decision makers. (1990, 4)

2.313 The Third Approach: Pragmatic Defenses

Pragmatic defenses are the most promising and have in one form or another often been offered in defense of various axiomatic conditions, but in the end they are also inadequate. Let me be clear what I am claiming here. It is not that *no* pragmatic defense of *any* axiom is *ever* adequate. Whether it is or not will depend on the details of the case. Rather, it is that it cannot be the case that *all* of the standard axioms can be given a pragmatic defense that does not depend upon particular preferences or relations among preferences for the person to whom the defense is offered. The satisfaction of some conditions – importantly including some conditions upon the agent's rationality – must be assumed to be in place before-hand.[53]

Since a pragmatic defense claims that we do better if our preferences and choice behavior conform to the axioms, the question I shall press is: Better *how* or in terms of *what*? I shall consider this in two stages. In the first, to illustrate some of the important points, I examine a paradigm of pragmatic argument in defense of an axiom, the 'money pump' argument in favor of Transitivity. In the second, I consider more generally what

[53] It is important also to realize that a pragmatic defense may turn out *not* to be a defense of the standard set of axioms. McClennen 1990 is, among other things, an extended pragmatic argument against the Independence axiom. (Or better, it is an extended pragmatic argument in favor of what McClennen calls 'resolute choice,' which, in certain circumstances, commits the resolute chooser to violations of Independence.)

pragmatic defenses of the axioms can be expected to do.

Suppose I have intransitive preferences over three kinds of fruit: I prefer apples to bananas, bananas to cantaloupes, and cantaloupes to apples. Suppose also that I have a cantaloupe. Since I would rather have a banana than a cantaloupe, you can induce me to pay you some small sum to exchange the cantaloupe for a banana. Once I have the banana, you can induce me to pay a small sum to exchange the banana for an apple. Once I have the apple, you can induce me to pay a small sum to exchange it for a cantaloupe. I'm back where I started, with the cantaloupe, except that I'm poorer. Even worse, if my preferences over fruit remain the same, you can repeat the cycle as many times as necessary to take all the money I have. (If you know those are my stable preferences and also how much money I have, you might even be well-advised to give me the cantaloupe if I don't already have one!) This is a version of the well-known money-pump argument for Transitivity.

The point of the story can be generalized. If my preferences are intransitive, I can be manipulated by others so that I inevitably end up worse off. Additionally, even without deliberate manipulation by others, sequences of events are possible in which, choosing on the basis of my intransitive preferences, I inevitably end up worse off.[54]

How compelling is this argument? Though it has considerable appeal, I am convinced that it is flawed. Its appeal derives from widely shared assumptions rather than from the strength of the argument itself.[55]

Consider again the case in which I have intransitive preferences over fruit and

[54] It is of course not essential to the structure of the argument that the losses made inevitable by my intransitive preferences be monetary.

[55] The argument I present, though developed independently, closely parallels the one in Hampton 1998, 244-247.

repeatedly pay to exchange a less preferred for a more preferred fruit. Why must I think that the result of such a sequence of exchanges is that I end up worse off?[56] If there is an answer, it appears that it would have to rest on the fact that I am assumed to have *transitive* preferences with regard to something else – in this case, with regard to quantities of money.

And this is not all, for I could have transitive preferences with regard to quantities of money and *still* think I am not made worse off by the series of exchanges. In particular, I would not think I had been made worse off if I transitively preferred less money to more. Then, I would regard the series of exchanges as improving my financial condition as well as, each time, replacing a less preferred with a more preferred fruit.[57]

But suppose I have more normal preferences with respect to money and transitively prefer more to less. Even so, this is not by itself enough to show that my intransitive preferences over fruits are in need of revision. I would have a set of preferences over quantities of money *and* a set of preferences over different fruits. The argument will only work if I must or think that I must regiment my preferences over fruit in terms of the preferences over money. But why must I do that? For all that has been said so far, the regimentation could go in the opposite direction – that is, if regimentation there must be, I could adjust my preferences over money so they didn't interfere with the fruit exchanges in which I wish to engage.

[56] A related point has been made by Loren Lomasky (in discussion). Why, he asks, could not the person with intransitive preferences argue that he is made better off by each exchange and therefore by all of them? After all, for any of the exchanges, had he not preferred making that exchange, he would not have done so. So, why can he not regard the money as well-spent (and the fruit well-exchanged), since he got something he preferred each time?

[57] Perhaps, if these were my preferences, I could be manipulated into worsening my financial condition by being required to accept a gift of a small sum of money with each fruit exchange!

So, what is needed for the money-pump argument to work is that, in addition to some set of intransitive preferences ranging over some domain, (a) the agent also has a set of transitive preferences ranging over some other domain, (b) that acting on the intransitive preferences insures that the transitive preferences invoked will be frustrated, and (c) that he considers it *more* important to satisfy the transitive preferences than those in the intransitive set. If those conditions hold, he will, if he can, adjust the intransitive preferences to make them transitive as well.[58]

But, this cannot work as a *general* argument for imposing transitivity upon one's preferences. For any given domain over which one's preferences intransitively range, the argument can provide a reason for imposing transitivity there only if there is some *other* domain in which one's preferences are already transitive (and with respect to which the other conditions are met). If there is no other domain within one's total preference set having the required characteristics, then no money-pump argument will work.[59]

[58] Strictly, still more conditions are needed. The agent would also need to be aware that he has an intransitive preference set and would need to think that the risk of being manipulated because of it was worth the trouble of trying to change it. (Suppose he had an intransitive preference cycle over a thousand elements. First, he would probably not be aware of it, and second, even if he were, might think it unlikely that anyone would be able to find and exploit the intransitivity.)

Also, if his judgment that it is more important to regiment the intransitive preferences by the transitive than *vice versa* is itself modeled as a preference – say, a preference in a domain ranging over options of preference-revision – then the preferences in *that* domain will also have to be transitive.

[59] It might be objected, without contesting the details of my line of argument, that my conclusion has little or no practical importance, given the preferences people actually have. It is, after all, extremely common for people to have transitive preferences for greater over lesser quantities of money and for them to take the fact that some course of action is sure to lose money, without any off-setting gains, as a decisive consideration against it. (Does the fact that I, with my intransitive preferences over kinds of fruit, am *pleased* to make each exchange count as an off-setting gain?) It might be said that as long as this (or some analogue) is true of the preferences people actually have, it hardly matters that money-pump arguments are not decisive for all possible preference profiles: It is enough if they are decisive for actual preference sets. They may still show that all of *us*, with the preferences we actually have, should regiment our preferences into transitive relations. Transitivity may be rationally, because pragmatically, binding upon us without being rationally binding for all logically possible agents.

As an objection, this is misconceived, for it concedes the point that the argument for Transitivity depends on the character of other preferences we happen to have. It just adds that we *do* happen to have the

Though I think this consideration of the money pump argument illustrates some important points, it is worth thinking further and more generally about pragmatic defenses of the axioms. A pragmatic defense of an axiom holds that we will, in some way, do better if we conform to it than if we do not. But how will we do better? So long as the theorist urging the pragmatic defense adheres to the decision-theoretic orthodoxy that there are no substantive requirements upon preferences, and so holds also that decision-theoretically rational choice does not presuppose that we hold particular preferences, the only option for the pragmatic defender would appear to be to claim that, by conforming to the axiom in question, we will do better in terms of our preferences.

But this is deeply problematic. To see why, recall some of the results already reached, namely, that satisfaction of all the axiomatic conditions is necessary to define a cardinal measure of preference satisfaction, and that the possibility of a cardinal measure is intimately connected with the existence of a complete ordering over options. Specifically, if there is a cardinal measure, then there must be a complete ordering. A slightly weaker claim was defended earlier about the converse relation: *for a finite mind*, if there is a complete ordering, then there is a cardinal measure. So, for a finite mind, there is a cardinal measure if and only if there is a complete ordering.

The problem we were considering is whether the person addressed by a pragmatic defense does better in terms of his preferences to conform to the axiom in question. But, *ex hypothesi*, his preferences do not satisfy all the axiomatic conditions upon preference sets. There are two possibilities. The first can be disposed of quickly. If a cardinal measure is needed to show that he does better in terms of his preferences, the argument

other preferences needed.

fails: it will not be the case that conformity to the axiom would better serve his preferences,[60] since no cardinal measure can be constructed unless all the axioms are satisfied.

The more interesting alternative is to deny that a cardinal measure is always needed to show what is better in terms of a set of preferences; at least sometimes, it is possible to show that one option is better than another in terms of a set of preferences without relying upon the existence of a cardinal measure. If this is assumed, then a pragmatic argument in defense of an axiom will succeed just in case conformity to the axiom is better in terms of the agent's preferences than non-conformity. That is, it will succeed when his preferences suffice to order the options of conformity to and non-conformity to the axiom and rank the first above the second.[61]

Suppose the pragmatic defense succeeds. This means the agent's preferences are such as to unambiguously order conformity over non-conformity to the axiom. There are two noteworthy implications of this fact. The first is that, since the agent addressed by the pragmatic defense does not already satisfy all the standard expected utility axioms and therefore does not completely order options in terms of his preferences, it is possible to have an ordering of preferences which is only partial but nonetheless sufficient for rational choice, at least within certain domains or over certain sets of options. This means, among other things, that the argument sketched earlier for the axioms from the

[60] Of course, this does not imply that he would do worse by conforming or equally well by not conforming. It just means that, so long as his preferences do not satisfy the axioms, there is no determinate answer.

[61] There are still of course ways in which a pragmatic defense of some axiom might fail. One is that the agent's preferences may be so disordered that none of his options in fact *are* ordered by his preferences. His preferences, so to speak, point in all directions and therefore not in any. Another is that though his preferences order some options, they do not order the options of conformity and non-conformity to the axiom. Still another is that, though conformity and non-conformity may be ordered by his preferences, conformity does not actually get ranked above non-comformity.

premise that rational choice presupposes maximizing must be mistaken. If any pragmatic argument at all works, then rational choice does not presuppose maximizing. Or, by contraposition, if rational choice presupposes maximizing, then no pragmatic defense of any of the axioms works. One cannot coherently hold both that rational choice presupposes maximizing and that pragmatic arguments can be given in defense of any of the axioms.

The second is that, in order for any pragmatic defense to succeed, the agent's preferences must already meet certain conditions. If any of those conditions are themselves to be defended as normative for choice, their defense must be conducted on other grounds. It is not possible in principle that a pragmatic defense can be given for all the axioms of standard decision theory,[62] or for that matter for all the axioms of whatever alternative to standard decision theory a particular pragmatic defender favors. Every pragmatic defense works, if it does, by relying upon the fact that the preferences of the agent to whom it is addressed already meet certain conditions.[63]

2.314 Can the Approaches be Combined?

In summary, none of the common approaches – not an appeal to what is

[62] Since every pragmatic argument rests on some assumptions about conditions met by the preferences of the agent to whom it is addressed, it might be that a pragmatic case can be made for conformity to *each* of the axioms without its being the case that a pragmatic argument can be made for *all* of them. The case for conformity to one would presuppose the satisfaction of certain conditions; the case for another would presuppose satisfaction of a different set of conditions, and so on.

[63] It might be thought that in principle a pragmatic defense either of an axiom or of the full set of axioms could be thoroughly dispositive. For suppose that there is some logically exhaustive way of characterizing sets of preferences, such that, say, all preference sets satisfy one of the sets of conditions, C_1, C_2 or C_3. Then, an argument for some axiom, A_1, might proceed to show that conformity to A_1 is pragmatically better relative to each of the sets of conditions the relevant preference set might satisfy. This apparent possibility, however, is an illusion, for it is surely logically possible that a preference set be such that it does not rank conformity and non-conformity to the axiom at all.

intuitively secure, not a coherentist defense and not a pragmatic defense – is capable of establishing that all of the axioms of standard expected utility theory are genuinely *requirements* upon the coherence of preference sets or, therefore, upon rational choice.

A natural question, then, is whether the different approaches can be combined in some way, so that what cannot be secured by any of the approaches operating individually is secured by their judicious joint application. It might be, for example, that some subset of the axioms is intuitively secure and that arguments of other kinds can be given for the remainder. It might even be thought that something like this is, in fact, somewhat inchoately at work in leading to the conviction, on the part of many decision theorists, that the full set of expected utility axioms is normative for choice. I do not know whether an argument of that sort can be adequately fleshed-out. Perhaps it can be. Still, it has not *been* done, so far as I know, especially not with careful attention to and distinction between what is taken to be intuitively secure, what is to be defended on the basis of broader coherentist considerations and what, given any previously defended conditions upon preference sets, is to be given a pragmatic defense. If it can be done, I would like to see the argument.

2.32 The Standing of the Axioms

In light of the foregoing, what can be said of the normative standing of the standard expected utility axioms? In their favor are both their mathematical elegance and tractability, plus the not inconsiderable attraction of the point that *if* one's preferences satisfied the axioms, it is very difficult to see how it could be denied that an action with a larger expected utility index is better in terms of one's preferences than any action with a smaller expected utility index.

A further attraction for some is perhaps better described as a motivation than as a reason for accepting the expected utility axioms. It derives from the thought that there must exist some procedure which amounts to an algorithm for resolving any decision problem with which one might be faced.[64] Applying the algorithm may be difficult in practice for any number of reasons, but, in principle, there is always a correct answer to be found, and it always *can* be found by correct application of the algorithm. For reasons that are not clear to me, some find the alternative that there are no algorithms in a given domain, but that we can sometimes identify and avoid mistakes or can sometimes find correct or better answers to questions posed within the domain, to be unacceptable. They are willing to tolerate any amount of practical difficulty in coming up with the correct answers, so long as they do not have to admit any theoretical indeterminacy in what the correct answers are nor that they lack methods for finding the correct answers.

What can be urged against the axioms is just the fact that, at the current stage of discussion, there is no adequate defense of the entire set of expected utility axioms. For an agent whose preferences do not satisfy the axioms, it cannot be simply a foregone conclusion that his choices, in light of his preferences, are less than rational. Of course, they may be less than rational, but pointing only to non-conformity to an axiom is not sufficient to establish the fact.

I think the most we can do at this point is to treat the claim that rational choice must proceed from a preference set that satisfies the expected utility axioms as an *hypothesis*. If the hypothesis is that it is both necessary and sufficient for a choice to be

[64] I take the existence of an algorithm for any decision problem to imply that there is some decision procedure which is sufficient to pick the best option, or one tied for best, out of any set of options that may be available in the given decision problem. A procedure that sometimes failed to do so, or sometimes failed to rank any option as best or tied for best, would not, in this sense, be an algorithm.

rational that it be based upon a preference set that satisfies the axioms,[65] then the hypothesis could in principle be tested in either of two ways. On one hand, we could try to find some case in which a choice based on such a preference set fails in some way to be rational. On the other, we could try to show that there are rational choices that are not based on such a preference set. For either kind of test, apparent failure will tend to count in favor of the hypothesis and apparent success against it.

For my purposes, I shall set the first kind of test aside[66] and concentrate entirely upon the second. Specifically, I shall argue that it is virtually certain that our preferences do not in fact satisfy all of the axiomatic conditions, but that this does not (nor does anything else) show that we do not or cannot make rational choices. Part of this I take to be obvious and uncontroversial: we can and at least sometimes do make rational choices, choices that are better than their alternatives in terms of our preferences and objectives. For that, I intend to offer no further argument than has already been presented. The other part, that we do not satisfy all of the decision-theoretic axiomatic conditions, requires more elaborate support. I shall begin with further consideration of the requirement that the elements of a preference set be completely ordered.

[65] A slight qualification is needed. It might be objected that the orthodox decision theorist would surely admit that a choice can be rational if based on a set of preferences ranging only over certain outcomes (and where only certain outcomes are relevant), when the preference set satisfied only the axioms of ordinal utility theory rather than the full set of expected utility axioms. This is true, but easily side-stepped. For the kind of case described, satisfaction of the expected utility axioms is sufficient but not necessary for what an orthodox decision theorist will recognize as a rational choice. For most cases, however, risky outcomes are relevant, and it is only those cases that concern me here, so the hypothesis can be expressed as the claim that, where risky outcomes are relevant, it is both necessary and sufficient for a choice to be rational that it be based upon a preference set that satisfies the expected utility axioms.

[66] It is difficult or impossible to find uncontroversial examples. Any proposal is likely to be met with the claim that the choice in question, though it may be counter-intuitive, is nevertheless rational.

2.321 Completeness and the Inscription Thesis

One of the axioms of expected utility theory is that the elements of a preference set must be completely ordered.[67] It must be possible to rank any element with respect to any other as preferred, dispreferred or indifferent to that other element. This is necessary in order to define a cardinal measure of preference-satisfaction and, consequently, for maximization in terms of the preference set to be well-defined.

If we are going to maintain that the Completeness condition is satisfied or may be satisfied for our actual sets of preferences, then a closer look at what is involved in having a preference is needed. For instance, we might suppose that an agent has a preference between two elements of his preference set, *A* and *B*, just when he has considered *A* and *B* together and either ranked one above the other or else ranked them as indifferent to one another. But if so, then it is clear that agents such as ourselves do not have complete orderings over our preference sets. There are innumerable pairs of items which are elements of our preference sets – that is, both of which enter into some preferential relation or other – but the members of which have not been compared to each other. I may prefer chicken over fish for dinner and Jones over Smith in the municipal election, but may never have considered whether I would prefer Jones's victory to chicken for dinner.

Generalizing a bit, if, in order to satisfy Completeness, I must explicitly compare each element in my preference set to each other, then, for three elements, *A*, *B* and *C*, I

[67] In what follows, I assume (as mentioned in note 46) that Completeness is not satisfied unless Transitivity is as well. Technically, however, the two requirements are independent. A preference set might be completely but not transitively, or transitively but not completely, ordered. Since I take it that there is little doubt that a coherent preference set must be transitive, I have chosen, except where it might make some difference to the argument, to speak only of Completeness.

need to perform three comparisons: A to B, A to C and B to C. For four elements, six comparisons are needed, for five elements, ten comparisons, and so on. Evidently, unless the number of elements is small, this is going to quickly get out of hand. For example, for 50 elements, 1225 comparisons would be needed.[68]

The point is even more obvious when we consider risky prospects. For between any two prospects in which, for the sake of argument, some outcome is assigned a probability of one hundred percent, such as having chicken for dinner (C) or Jones's electoral victory (J), there are infinitely many gambles, corresponding to the infinitely many possible values of p (with $0 < p < 1$) in $[C, p; J, 1 - p]$. Before, for some finite number of elements, the task of comparing them, if the number of elements is large, was (merely) forbiddingly difficult. Here, for Completeness and Closure both to be satisfied, each of the infinitely many gambles between chicken for dinner and Jones's victory must also be preferentially ranked with respect to every other element of the preference set (including every other gamble over any other elements of the preference set!). So, if preferential ranking presupposes explicit comparison, the task is, for finite minds (over finite periods of time), strictly impossible.[69]

Taken together, these facts imply the following disjunction: Either we do not (and *cannot*) satisfy the Completeness condition, or else, preferential ranking does not presuppose explicit comparison. Accordingly, if we assume that it is possible for us to satisfy the Completeness condition, we must also assume that preferential ranking of the elements of a preference set is possible without explicit comparisons. It can be true that

[68] In general, for n elements, the number of comparisons needed is equal to the sum of the integers from zero to n - 1.

[69] I assume that there is some minimum, non-zero, time required to perform an explicit comparison.

an agent has some preference ordering over prospects that she has never considered together or, for that matter, has never considered at all.[70]

Let us see what this implies. If we assume that some agent's preferences satisfy the Completeness condition (together with the other axioms), the most obvious and most important implication for my purposes is something that I shall call *the inscription thesis*. The inscription thesis holds that the preferences or preferential relations involved or expressed in explicit comparisons have an underlying structure, not necessarily attended to but which is nonetheless present within those preferential relations, and which is sufficient to determine the remaining preferential relations between all the elements of the preference set.[71] The preferences involved in explicit comparisons have, inscribed within them, so to speak, all the preferential relations among all the elements of the agent's preference set. The preferential relations of the options explicitly considered embody already strengths or degrees or weights that can be compared to one another. Some limited number of explicit comparisons has been performed and preferential rankings between the items compared have been established, but somehow there can (truly) be ascribed to the agent a complete ordering over all the elements of her preference set, including those that have never been explicitly compared.

The inscription thesis appears to me very doubtful, and in what follows, I shall try to cast further doubt upon it. But before doing that, I will address two lines of defense

[70] She may never have considered either of a pair of prospects at all because the presence of each as elements in her preference set is guaranteed by the Closure axiom. She may rank A over B and C over D, by virtue of having explicitly compared them, but never have considered, for specific values of p and p^*, either $[A, p; B, 1 - p]$ or $[C, p^*; D, 1 - p^*]$. Nonetheless, for Completeness to be satisfied, it must be true that she has a preferential ranking between those two gambles.

[71] By "sufficient to determine," I mean that the underlying structure has features such that there is a unique answer as to what the further preferential relations are.

that might be offered.

2.3211 Two Defenses of the Inscription Thesis

A defense of the inscription thesis might be grounded in the claim that if the agent were presented with a choice between any pair of the elements in her preference set, including among those elements all of the gambles that can be defined over other elements, then she would make some choice or other.[72] In this form, the proposal faces crippling objections. One is that unless determinism is true (or true for the conditions under which she is supposed to make the choice), it may simply be false that she would make some definite choice. She would make some choice or other, but there may be nothing about her or her situation that settles which choice she would make. And, even if determinism does hold for the conditions under which she is supposed to be choosing, it may be true that she would make some definite choice between the options presented to her, but it does not follow either that she prefers that option to the other or that she is at least indifferent between them, unless we assume, what we have already rejected, some version of a revealed preference theory. The fact that she makes a certain choice is not sufficient to show that she has at least a weak preference for what is chosen over what is not, for it may be that the counterfactual, 'if she were presented with a choice between A and B, she would select A,' is true, but true in virtue of some feature of her situation other than her preferences.

So, the suggestion has to be amended to say that if she were presented with any pair of elements in her preference set (whether for choice or not), she would have some

[72] Dawes (1988, 154-155) suggests something like this in defense of Completeness.

preferential ranking between them. Now, it does not seem obvious to me that this is true, but even if it is true, that is still not sufficient to underwrite the inscription thesis. For there is still the possibility that, if presented with such a pair, she would then *form* a preferential ranking between them – perhaps some definite preferential ranking, not just some preferential ranking or other – but that the preference she then forms is not determined by her pre-existing set of preferences.[73]

But suppose we avoid this possibility as well and assert that if she were presented with any pair of elements in her preference set (whether for choice or not), she would have some preferential ranking between them that is determined by her preference set as it was before she was presented with the alternatives. *Perhaps* this is so, but it seems exactly as doubtful as the inscription thesis itself, for the simple reason that it is equivalent to the inscription thesis: *This* counterfactual will be true just in case the inscription thesis is true, and false otherwise; hence, its truth cannot provide the inscription thesis with any independent support.

The other approach to defending the inscription thesis can be treated more briefly. It appeals to the techniques for constructing a cardinal utility function and points out that, for any elements of a preference set that can be located with a cardinal measure along a real-numbered scale, the preferential relations in which they stand to one another, including the preferential relations between all gambles defined over the elements of the preference set, can be derived. The idea is that only a few fixed points are needed rather than an infinite set of comparisons. The rest of the preferential relations are functions of the few fixed points. But as a defense of the inscription thesis, this is confused, because

[73] It is, if the counterfactual is true, presumably determined by *something*, but what determines the preferential ranking need not be her preference set, or her preference set alone.

the argument is circular. A cardinal utility function can only be defined for a preference set if the expected utility axioms, including Completeness, are satisfied. But it was the apparent fact that Completeness might not be satisfied by actual preference sets that was the rationale for interpreting Completeness in terms of the inscription thesis. Of course, *if* Completeness and the other axioms are satisfied, the inscription thesis is true (for any finite mind), but that yields no assurance that Completeness and the other axioms *are* satisfied.

So far, I have examined the only two arguments I know for the truth of the inscription thesis and found both wanting. Still, it *might* be true. Or more precisely, it might be that the inscription thesis is true of the preference sets of at least some of us. So, what I will do at this point is to turn to presenting a series of considerations against the truth of the inscription thesis. I do not know of any direct way of demonstrating that the inscription thesis is false – false, that is, with respect to the preference sets we have – but I think it can be thoroughly undermined: it can be shown that it is very unlikely to be true and thus, that it is not reasonable to believe that all of the weightings or degrees of strength needed for the truth of the inscription thesis are actually present in our preferences.

2.3212 Undermining the Inscription Thesis

I shall look at three kinds of considerations[74] which are aimed at showing that it is

[74] There is another kind of argument, to which I have already alluded, for the incompleteness of most persons' preference sets. This consists of the considerable empirical evidence that people regularly and systematically violate the axioms of expected utility theory. (See, e.g., Dawes 1988 and Kahneman and Tversky 1990 – which represent only a small sampling from a very large literature.)

This kind of evidence, however, is widely known and has not prevented people from thinking that implausible to think that we *can* satisfy the axiomatic conditions. The first two have to do with *novelty*, with either new objects of preference or with previously unconsidered decision problems, while the third has to do with *uncertainty* (and its relation to probability). There is some overlap between these, and it will not be possible to keep them completely separate, but that fact has certain advantages for my thesis, for it implies that the considerations cannot be answered in isolation. An adequate answer to one will have to address the others as well, so far as they overlap with it.

2.32121 Novel Objects of Preference

The first issue to consider is how to understand what happens to a preference set when some new element is introduced, for it is an important fact about preference sets that they have histories. The relatively simple preference sets of children become, as the children mature, more complex and come to include elements about which their younger selves would have had no preferences. This may happen in many ways, but what is important here is the extent to which this is a matter of novel experience introducing an agent to something which he did not previously rank preferentially at all. Whether it be a matter of new tastes or sensations, new activities, or new dimensions of concern or interest, they must somehow be integrated into and thereby alter the agent's pre-existing preference set. A great deal of this kind of change must occur in the course of a normal agent's life.

satisfying the axioms represented an appropriate ideal. My intentions are more radical – not to argue that we fall short of satisfying the axioms and therefore should try harder, exercise more care or the like, but instead that there are deep reasons for thinking that satisfying the axioms is not something that we can do and therefore is not, for beings like us, an appropriate ideal – which of course is not to say that there are no standards of rational choice appropriate to us and in light of which we *should* try harder, exercise more care and the like.

If the inscription thesis is true of an agent who must integrate some novel element into his preference set, and if it remains true after the new item is integrated, then the preferential relation between that item and everything already a part of his preference set – its relative weight or importance – must somehow be *inscribed* into his preference set by virtue of his coming to preferentially rank it. How is this inscription process supposed to work?

Consider the following schematic illustration. An agent prefers *A* to *B* and *B* to *C*. Such an ordinal ranking is of course not sufficient to insure that this is a complete ranking, even within this limited range. For that, we must suppose also that there is some definite gamble between *A* and *C* such that [*A*, *p*; *C*, 1 - *p*] is indifferent to *B*. But let us grant that. What happens when the agent considers some previously unranked option, *D*?

Suppose the agent considers the relation of *D* both to *A* and to *C* and is sure that *D* falls somewhere in the range (exclusive of the endpoints) between *A* and *C*. Thus, *A* is preferred to *D* which is preferred to *C*. For *D* to be fully integrated into a complete preference ordering, though, he must establish at what gamble between *A* and *C* he would be indifferent between the gamble and getting *D*. Moreover, to avoid introducing intransitivities into his preference set, he must get this *exactly* right. It is not good enough to conclude that he is indifferent between *D* and the gamble, [*A*, .6; *C*, .4], if it would be more accurate to say he is indifferent between *D* and the gamble, [*A*, .599; *C*, .401]. In particular, the gamble between *A* and *C* that he accepts as indifferent to *D* must stand in exactly the right relation to the corresponding gamble with respect to *B* – which he has not considered in establishing the ranking!

It does not much matter here how likely it is for the agent to assign exactly the right value to the new object of preference, provided only that it is less than certain.[75] Even if he will very likely get each one right, the fact that the preference sets of adults have come to be as they are through the addition and ranking of *many* new elements insures that it is enormously unlikely that *all* of the new rankings are exactly correct.[76] And if any such ranking is not correct, there will be intransitivities and therefore failures to completely and unambiguously order his options.[77]

2.32122 Novel Decision Problems

Novelty is an issue for the completeness of preference sets in another way, connected with previously unconsidered decision problems. It is a familiar fact that in the normal course of events we are faced with decision problems that we have not faced nor even considered before. Some choice must be made from among a set of options, when the agent has never compared all of them to one another.[78]

To adapt an earlier example, suppose Caroline knows she would prefer chicken to fish for dinner and Jones to Smith in the municipal election. But she has never compared chicken for dinner to Jones's victory. Still less has she ever compared her actual options,

[75] It does not help to claim that there is no sense to 'getting it right' that is independent of the actual ranking assigned. Apart from other problems, such as its close kinship with a revealed preference view, there is one obvious way of getting the ranking wrong: it may turn out to be probabilistically incoherent with other rankings one would assign. If so, they cannot all be correct.

[76] Suppose there is a ninety-nine percent chance that each new object of preference will be ranked correctly. Then there is only about a thirty-seven percent chance that all of one hundred new objects of preference will be ranked correctly. For larger numbers (or a smaller chance of getting each one right), the chances are of course even less.

[77] Strictly, the argument for this presupposes that at least four non-compound elements of a preference set are being ranked *vis-à-vis* one another.

[78] Here, I am restricting myself to options that are *currently*, in advance of the choice, part of the agent's preference set. By this I mean that each of the options can be defined in terms of elements of her preference set that are present because, at some point, there was an explicit comparison and ranking with respect to at least one other element.

a high probability of chicken for dinner to a slightly increased chance that Jones will win. But we can suppose that circumstances force the choice upon her. How is she to make a decision between options she has never before compared?

If the inscription thesis is true of Caroline, the answer must somehow be there, present to be elicited, in her preferences. She must really, albeit not yet explicitly, prefer the high probability of the chicken dinner to the slightly increased probability of Jones's victory, or *vice versa* or else be indifferent between them.

Now, there are two sorts of cases where it seems that it may be true that the answers really are present to be found in her pre-existing preferences. First, Caroline may immediately know which she prefers as soon as she realizes that she is facing the choice. If she were compelled to choose between having her thumb smashed by a hammer and having chicken for dinner, we would be surprised if she hesitated to answer, even if she had never explicitly compared those options before. If the choice between the chicken dinner and Jones's victory were like that, it seems reasonable to say that the answer was already present in her preferences. Second, she may not immediately know how she ranks the options, but, in reflecting upon them, she comes to some definite conclusion, perhaps by taking into account others among her preferences and how they will be affected respectively by the chicken dinner and Jones's victory. She may have no sense that she is doing anything other than more richly articulating, and thereby bringing to the surface, preferences she already has. If this is what is going on, it may well be that she is finding an answer that was already present in the structure of her preferences.

If the novel decision problems Caroline faces are all of one of the foregoing two sorts – where she immediately or upon reflection realizes how her preferences bear upon the decision problem – it might be reasonable for her to suppose, at least so far as this is the only relevant issue, that the inscription thesis is true of her. But there is another type of case in which it seems more doubtful that she is only discovering something already true of her preferences.

Suppose that when Caroline is faced with the choice between the relevant probabilities of chicken for dinner and of Jones's victory, she does not immediately know how she ranks them (so it is not a case of the first sort), but also that further reflection does not yield any answer (so it does not appear to be of the second sort, either). Since we are supposing this is a forced choice, she will of course *select* one or the other, but may still say that her selection is not a matter of *preferring* one to the other nor is it a matter of being indifferent between them. She is not confident that her selection has the best expectation of serving her preferences nor that it can be expected to do as well as the alternative.

Now, in a case like this, especially if Caroline has had ample time to elicit her preferences between the options, I think we should take her word for it: She has not succeeded in eliciting a preferential ranking between her options because it was not there to be found.

But it is of course possible to maintain, despite Caroline's actual non-success in finding it, that the answer is still somehow present in her preferences. It is not obvious that this helps for two reasons. First, it may raise a parallel question on a different level, namely, what to do in selecting between a pair of options when, under the circumstances, a preferential ranking for them cannot be established (or, perhaps, discovered). Should she flip a coin, substitute an easier problem (e.g., act as she would if she thought her choice would completely determine whether she got chicken for dinner or whether Jones got elected), act on impulse or what? Surely her response might be that she cannot establish a preferential ranking for *those* options, either.

But there is a second and more interesting level of response. No one ever has unlimited time or cognitive capacity to investigate the content of her own preferences. Accordingly, since it is not being assumed that all preferences must be either conscious or instantly available to reflection, anyone may, in a given case, be incapable of determining how some option stands with respect to her preferences. Nonetheless, we typically assume that, under good conditions,[79] agents are reliable (not infallible) in determining what their preferences are. Suppose that Caroline is not especially rushed in coming to a conclusion and that there are no obvious interfering factors, but that she reports that she is confident that the answer is not there to be found – that is, that the answer is not already present in her preferences. She simply does not know how to rank the options. Surely, this is possible.

But there is a dilemma here for friends of the inscription thesis. On one hand, Caroline says that her preferences are not sufficient to rank her options, and there is no special reason to doubt her reliability. But if the inscription thesis is true of her, she must be mistaken. On the other hand, we have no *better* warrant for accepting her reliability in the kind of case in which she reports, after reflection and under good conditions, that her preferences *do* order her options. Just as she could be mistaken in thinking that her pre-existing preferences do not order a pair of options, she could also be mistaken in thinking that her pre-existing preferences *do* order a pair of options. Just as she might have overlooked or failed to attend in the right way to some feature of her preferences that

[79] This is vague, and I shall not try to spell it out, but 'good conditions' are meant to rule out various circumstances, such as distractions or tiredness, that can be expected to interfere with or distort judgment.

would succeed in ordering a pair of options, she might have overlooked or failed to attend in the right way to some feature of her situation, extraneous to her preferences, that determined the apparently successful ordering she reported. If her judgment in one case is doubtful, and therefore not sufficient to show that her preferences fail to completely order her options, it appears that her judgment in the other case is equally doubtful and therefore not sufficient to show that her preferences in *that* case completely ordered her options. What is gained on one hand is lost on the other.

It seems to me that issues such as Caroline faces here, though perhaps not common,[80] may affect many decisions. By this, I do not mean of course that many of us are faced with deciding between chicken and the victory of a favored candidate, but that almost any of us can be placed by circumstances in a position in which we do not know how to rank the options we face and in which further examination of our preferences leads no closer to an unambiguous ranking.[81] If this is both correct and correctly represents the actual extent of order in our preference sets, then our preferences do not completely order our options.[82]

2.32123 Uncertainty

There is a third reason for the claim that our preferences are incomplete. To this

[80] It is difficult to be sure just how common they are, since they may often be present when there is insufficient time, before a decision must be made, to identify them and reliably rule out alternative explanations.

[81] That this strikes me as plausible may only show that *my* preferences are not complete!

[82] The plausibility of this conclusion is reinforced by the fact that, for someone in a position like Caroline's, a very natural response – if a decision need not be made immediately – is to turn from considering which of her options she *does* prefer to considering which she *should* prefer. She may of course find no answer to that question, either, but the fact that she raises it and hopes to find an answer implies that she does not think her existing preferences provide everything needed to make a decision.

point, I have for the most part spoken as if the probabilities to be assigned to outcomes were unproblematically available. Further, it is important, if the axiomatic conditions are to be satisfied, that these probabilities be entirely definite point-probabilities,[83] so I have in effect assumed as well that point-probabilities are unproblematically available. To bring out what this involves, we need to look at the contrast case, at the alternative to the availability of point-probabilities.

The conventional way to do that is to draw a distinction between *risk* and *uncertainty*.[84] Conceptualizing a situation as one of risk involves a particular characterization of an agent's ignorance or knowledge of the future. A person making a decision under conditions of risk may not know what the outcome of his decision will be but knows exactly the probabilities attaching to the different possible outcomes. For instance, a person considering playing Russian roulette knows (assuming the gun is working properly) that if he pulls the trigger he has a one-in-six chance of getting a bullet in the head and a five-in-six chance of not getting a bullet in the head. Certainty about what the future holds is just a limiting case of risk, one in which a single (non-compound) outcome has a probability of one hundred percent.

The polar opposite case of ignorance about the future can be illustrated in this way. Suppose the person considering playing Russian roulette does not know how many

[83] Consider three elements of a preference set, A, B and C, and suppose that A is preferred to B and that B is preferred to C. Assume further that there is some definite value, p, such that $[A, p; C, 1 - p]$ is indifferent to B. Then, if $1 \geq p^* > 0$, $[A, p^*; C, 1 - p^*]$ must be preferred to C, and if $1 \geq p^{**} > 0$, $[B, p^{**}; C, 1 - p^{**}]$ must be preferred to C. But unless p^* and p^{**} have entirely definite values, it will not always be possible to compare $[A, p^*; C, 1 - p^*]$ and $[B, p^{**}; C, 1 - p^{**}]$. It will be consistent with the suppositions that the first would be preferred to the second, that the second would be preferred to the first or that they are indifferent to one another. But if that is the case, Completeness will not be satisfied.

[84] I have written elsewhere (1994) on problems of choice under conditions of uncertainty and reached no definite conclusion save that it is a hard problem and that the most plausible proposal for converting it into an easy problem, the one to be outlined here, is by no means rationally compelling.

chambers the gun contains or how many chambers are loaded. There may be any number of chambers, *n*, and any number of them, from zero up to *n*, may be loaded. Then, he would have *no* definite probabilities assignable to the possible outcomes of getting or not getting a bullet in the head. He is completely uncertain, with respect to those alternatives, what the future will hold if he pulls the trigger.

Once both risk and uncertainty have been characterized, we can of course imagine any number of intermediate cases of partial uncertainty. For instance, the potential Russian roulette player may know that the gun contains either six or nine chambers and that either one or two chambers are loaded, so, assuming he doesn't want a bullet in his head, he has at worst a one-in-three chance of getting a bullet in the head and at best a one-in-nine chance. He would know that the probability of getting a bullet in the head, if he pulls the trigger, can be represented by some value in the set, {1/9, 1/6, 2/9, 1/3}. Further elaboration of the example could yield probabilities equivalent to some unspecified value in the closed interval between one-ninth and one-third ($1/9 \leq p \leq 1/3$) – or, for that matter, within any other interval of probability values. Or – not so readily exemplified by ringing changes on possible arrangements for Russian roulette – there may only be ordinal probability information available to the agent: he may know that one outcome is more likely than (or about as likely as, much more likely than, etc.) another.[85]

There are at least three important points here. The first two are fairly obvious; the third requires a bit of elaboration. The first is that most ordinary reasoning about

[85] Ordinal probability information is not reducible to some interval probability. '*A* is more likely than *B*' is not equivalent to the probability of *A* being equal to some value in the interval, $.5 < p \leq 1$ (and the probability of *B* being equal to some complementary value in the interval, $0 \leq p < .5$), because there may be some alternative or set of alternatives to *both A* and *B*, the probability of which is unknown or only partially known.

probabilities bearing upon decisions to be made in fact involves at least partial uncertainty rather than risk. Our probability-judgments do not usually assign well-defined point-probabilities; instead, they usually take the form of assigning approximate values or else are simply ordinal. Moreover, there is reason to think that this is not an *eliminable* feature of our reasoning about the unknown future, something that we could replace with point-probabilities if we were more careful or more assiduous in gathering evidence. Apart from other difficulties,[86] this is clear because one of the sources of partial or complete uncertainty in probability-judgments is the *known* possibility of *unknown ignorance*. There may be some possibility relevant to a prospective decision which is not recognized at all, and which is therefore not assigned any probability value.[87]

The second point is that if genuine and irreducible uncertainty (whether complete or not), as distinct from risk, characterizes what an agent knows about the future, there is no well-defined sense to maximizing.[88] The way in which decision-making under risk is assimilated to maximizing is to replace, as the appropriate maximand, actual utility with

[86] One is the fact that greater care or additional effort in gathering evidence may themselves demand resources, especially in the form of time, that are not available at the time a decision must be made.

[87] Abstractly, a case of unknown ignorance can be described in this way: Suppose an agent considers the probability to be assigned to each of a pair of outcomes, *A* and *B*. Suppose also that he assigns fully definite point-probabilities to each. That may be a mistake, for there may be some other outcome, *C*, which he has not considered and to which he has assigned no probability. Since the unconsidered possibility, *C*, may make a difference to what he would decide (had he considered it), the probabilities assigned to *A* and *B* can at best be sufficient basis for judging (say) that *A* is more likely than *B*, but not for judging that the probabilities assigned to *A* and *B* give the correct factors to be employed in weighting the respective utilities of the two.

For an example, consider an agent estimating her potential liabilities under the terms of a contract. She may, having taken all reasonable steps, conclude that her maximum liability is a certain sum and that the advantages she expects to derive can be represented by some different sum. She may assign probabilities to the various events that condition these gains or liabilities and conclude that she should sign the contract. However, it may be that her actual potential liability, due to some inadequately understood provision of the contract, is much greater and that, had she been aware of it, she would have concluded instead that she should not sign the contract.

[88] This will be qualified somewhat below.

expected utility, where the expected utility of an outcome is its actual utility discounted by its probability, and the expected utility of an option is the sum of the expected utilities of the various possible outcomes of selecting that option. But when point-probabilities to assign to the outcomes are not available, it is unclear by what the utilities of the various outcomes are to be weighted or discounted in order to determine which option has the greatest expected utility.

The third point requires a bit more background. There is a way of extending standard expected utility theory to cover cases of uncertainty.[89] The result of the extension, *subjective expected utility theory*, requires a strengthening of the axioms, especially Completeness, so that what is necessary to satisfy Completeness includes a complete ordering over not only risky prospects but also over *uncertain prospects*, where an uncertain prospect is one that may include partial or complete uncertainty about the probability to be assigned to its constituents.[90] Then, the relevant maximand is *subjective expected utility* – that is, expected utility ranging over uncertain prospects and based on whatever probability information or beliefs the agent has (his subjective probabilities, as these are called).[91] Once we admit uncertain prospects as elements to be ordered in a

[89] Useful discussion may be found in Luce and Raiffa 1985, Chapter 13.

[90] To smoothly extend standard expected utility theory to cover these cases, it should be assumed that certain and risky prospects are special cases of uncertain prospects. Certainty will be conceived as varying from zero, or complete uncertainty, to one, or complete certainty. It is an interesting question whether the variation in uncertainty should be conceived as admitting infinitely many values or degrees (whether continuously or not) or whether there is some finite set of degrees of uncertainty in the range. I think that we cannot limit ourselves to finitely many possible degrees of uncertainty. For suppose that there are at least two degrees of partial uncertainty. Suppose that one of them holds when all the probability information available is that one option, *A*, is more likely than another, *B*, and that the other holds when *A* is much more likely than *B*. Could we not then construct a further uncertain choice in which it is, say, completely uncertain whether the first or the second is true, and will not that further choice have to have some degree of uncertainty not to be identified with either of the other two? Now, with a third degree of uncertainty, we can repeat the argument, comparing it with either the first or the second, to get a fourth, and so on.

[91] There are of course further requirements attendant upon the inclusion of uncertain prospects and

preference set, then, if the preference set satisfies the subjective expected utility axioms, it can be shown that any outcome for which one has only limited probability information can be treated as indifferent to some gamble from a reference set that can be expressed in terms of point-probabilities.[92]

So if, for instance, an agent is faced with partial uncertainty in a choice in which he believes a given action will lead to either *A* or *B* as an outcome and believes that *A* is more likely than *B*, there will have to be, by the appropriately strengthened versions of Probabilistic Continuity and Strong Independence, some gamble between *A* and *B* such that the agent would be indifferent between facing it and the corresponding uncertain prospect. That is, there must be some value of *p* such that the agent would be indifferent between being offered the gamble, $[A, p; B, 1 - p]$, and being offered the uncertain prospect between *A* and *B* described above.

Once we have gone this far, it is clear that choices under uncertainty, whether partial or complete, can be assimilated to maximizing. Even if, as claimed above, we cannot eliminate uncertainty about the future, we are able, by finding gambles that can be expressed in terms of point-probabilities that are indifferent to uncertain prospects, to say by what values the associated outcomes are to be weighted in decision-making. This is sufficient to show that if the agent's preferences and subjective probabilities satisfy certain conditions, then we can make sense of saying that there is some quantity maximized in rational choice.

subjective probabilities. They must satisfy the axioms of expected utility theory (or appropriately strengthened versions of them), and, in particular, the subjective probabilities must also conform to the usual rules for combining probabilities.

[92] This conclusion about limited probability information extends to what might be termed the maximal case of limited probability information: the case in which one has *no* probability information – that is, to choice under conditions of complete uncertainty.

This is not to say, however, that it is plausible that anyone actually does satisfy the conditions. For it is a significant fact that the conditions to be satisfied are more demanding than those of standard expected utility theory. Since it is vastly implausible that our preferences satisfy the conditions of standard expected utility theory, it is even less plausible that they satisfy the more demanding conditions of subjective expected utility theory. To put it differently, if standard expected utility theory is to be applicable to an agent's preferences, then those preferences must satisfy the inscription thesis: somehow, the strengths or weightings of all risky prospects, including ones never explicitly considered but constructed from elements of his preference set, must be present in his preferences and their relations to one another. For subjective expected utility theory, what must be 'inscribed' in the agent's preferences must include not only all of those, but must include in addition relative strengths or weightings for all uncertain prospects that can be constructed from elements of his preference set. For any pair of non-compound risky prospects, *A* and *B*, there can be constructed, in addition to all of the infinitely many gambles possible between *A* and *B*, at least one uncertain prospect,[93] one in which the agent is completely uncertain whether she gets *A* or *B* (designate this as [*A* ○ *B*][94]). This uncertain prospect will have to be assigned a utility index such that it is indifferent to some gamble between *A* and *B*. But [*A* ○ *B*] will also enter, as an element, into the construction of further risky and uncertain prospects, for example, into gambles

[93] In fact, there will be at minimum several uncertain prospects, reflecting differing degrees of uncertainty, that are not equivalent to some (merely) risky prospect, e.g., in which *A* is more likely than *B*, in which *A* is much more likely than *B*, in which *A* is about as likely as *B*, and possibly more. In principle, if we take seriously the idea that uncertainty can vary among infinitely many values (see note 90), there will be infinitely many uncertain prospects that can be constructed using *A* and *B* as elements. So, the problem of integrating uncertain prospects into an agent's preference set is actually harder than outlined in the text.

[94] I adopt "○ " simply to represent some operation of concatenation between prospects, such that the outcome of the operation is uncertain (as distinct from risky).

against risky prospects, such as $[[A \circ B], p; C, 1 - p]$, into compound uncertain prospects, such as $[[A \circ B] \circ [C \circ D]]$, and so on. Since the inscription thesis, adjusted to accommodate uncertain prospects, requires more than the version adapted to merely risky prospects, it must be less likely that it is true of an agent.[95]

2.3213 Incompleteness

I have been arguing that three kinds of considerations – related to novel elements in a preference set, to novel decision problems and to uncertainty – show that our actual preferences do not satisfy the inscription thesis.[96]

[95] It is implausible, then, that an agent's preferences with respect to uncertain prospects satisfy the relevant version of the inscription thesis, but there is also a consideration that calls into question whether maximizing subjective expected utility is rationally required, even for the imaginary agent whose preferences *do* satisfy the relevant conditions. In terms of subjective expected utility theory, we can understand what would have to be true of an agent's preferences in order to characterize his choices under uncertainty as maximizing some quantity, namely, subjective expected utility.

But the appeal of maximizing lies in the thought that the maximizer does better in some relevant sense than an otherwise similar agent who does not maximize. In the face of certain outcomes, the utility maximizer does better than the non-maximizer by achieving outcomes that he at least weakly prefers to all others, while the non-maximizer does not. In the face of risk, the expected utility maximizer does better on average than the non-maximizer. In both cases, we have some understanding of the sense in which the maximizer does better than a non-maximizer with the same preferences in the same situation. But, in the face of uncertainty, the subjective expected utility maximizer does better .... how? Only in terms of a metric constructed so as to have a consistent way of saying what it is to do better under uncertainty. What is not clear is that he does better in any other, intuitively acceptable, sense. He need not do better as things actually turn out, nor on average, nor at avoiding disasters, nor at achieving benefits nor in any other independently specifiable way. Moreover, it is not just that the maximizer of subjective expected utility *may* fail to do better in any independently specifiable way, but also that there cannot be a sound argument that he would do better in some other way, for, if there were, we would not be dealing with genuine uncertainty.

What this means, I think, is that here it is the requirements for mathematically representing a quantity that can be maximized that is driving the argument rather than the plausibility of maximizing as a rational requirement. Put differently, since, when uncertainty enters the picture, maximizing has no well-defined independent sense, there is no warrant for supposing that it is a rational requirement. We can *give* it a sense along the lines of subjective expected utility theory, but even for agents, if there are any, whose preferences satisfy the subjective expected utility axioms, we cannot provide any further argument that they will do better to maximize subjective expected utility rather than adopt some other course. They will have only the Pickwickian consolation that they will do better at maximizing subjective utility.

[96] An additional powerful reason for thinking that our preferences do not fully order our options derives from the prevalence of unconscious motivation and particularly from unconscious biases that may shape what options are considered. If some options are systematically prevented from coming into

These considerations have two important features in common. The first is that they appeal to cognitive limitations rather than to rational defects. What I mean can be brought out as follows. It is not possible to sharply distinguish cognitive limitations from rational defects, since anything that may be called a rational defect – e.g., the disposition to endorse arguments that have the form of denying the antecedent (that is, $P \rightarrow Q; \sim P; \therefore \sim Q$) – can also be represented as a possibly severe case of cognitive limitation. Nonetheless, we do, in a rough and ready way, distinguish the two, attributing some mistakes to ignorance or inability and others to defects in reasoning. Roughly, we say that there are rational defects when we think the inference or action in question is one that the person (rationally) should, given his knowledge and abilities, have gotten right, but did not. That, of course, depends on what we take his knowledge and abilities to have been – that is, upon what his cognitive limitations are – so drawing the distinction presupposes that we already know something about which mistakes result from cognitive limitations and which from rational defects. Though we cannot make the distinction perfectly sharp, we still have excellent reason for drawing it, for, on one hand, it would be impossibly demanding to treat all mistakes as the product of rational defects, and, on the other, intolerably lax to excuse all mistakes as due to cognitive limitations. The fuzziness of the distinction does not matter here, so long as it is granted that the kinds of considerations I urge against the truth of the inscription thesis need not involve rational defects.[97]

consideration, then their preferential ordering with respect to other options comes into question. Be that as it may, I do not need to rely upon any such arguments.

[97] If it is objected that the considerations I have urged *do* rely upon the presence of rational defects, it is incumbent upon the objector to specify what those are, and importantly, to do so without begging the question – specifically, without relying upon the assumption that any failure to satisfy the inscription thesis

To return, two of the considerations, the first and third, are in essence arguments from finitude and rely upon the implausibility of the claim that we can correctly and consistently solve problems the solutions of which require unlimited precision. The second is more complex. It relies upon three points, first, upon the plausibility of the claim that anyone can be faced with decision problems such that, even under good conditions, she cannot determine which of her options best serves her preferences, second, that this is good evidence that her preferences do *not* completely order her options and therefore do not satisfy the inscription thesis, and third, that any argument casting doubt on the quality of the evidence and therefore on the conclusion would equally, though in a different way, cast doubt on the genuineness of the cases in which her preferences *appear* to order her options.

None of these three arguments assumes that we are rationally at fault for the incompleteness of our preferences.[98] We would be at fault only if there were some procedure available to us for preference acquisition and modification that would non-accidentally result in the completeness of our preference sets. Since we have no such procedure, the incompleteness of our preferences, and therefore their failure to satisfy the inscription thesis, must be ascribed to cognitive limitations.

There is a second important feature which follows from the first. Given that our preferences do not satisfy the inscription thesis, there is no obvious way of *coming* to satisfy it that is not subject to the same problems. If anything, the problems are

is *ipso facto* sufficient evidence for the diagnosis of a rational defect.

[98] This is not, of course, to deny that failures of rationality may be operative in generating sets of preference that do not completely and unambiguously order options. However, irrationality was not relied upon in the arguments presented, so those arguments amount to a case that even if no rational defects are involved in preference formation or revision, we have strong reason to think that resulting sets of preferences will fail to satisfy the inscription thesis.

compounded in the absence of the assumption that the preference set to be revised in order to satisfy the inscription thesis is already complete and transitive. When the problem is only, at a given stage, to integrate a single new object of preference into the elements of a preference set that is already supposed to be complete and transitive, its relations and weighting with respect to other elements of the set must, indeed, be gotten exactly right to avoid introducing intransitivities. However, provided that there is some way to detect that an initial assignment of relative weight to the new element is not correct, only *its* weight needs to be adjusted. But when the preference set is not supposed already to be complete and transitive, it is not evident either where to begin or where to stop. In particular, imposing transitivity upon some subset of the elements in a preference set may introduce intransitivities in other, over-lapping, subsets; rectifying those may introduce yet further intransitivities, and so on.[99] If intransitivities are present, they *may* be uncovered by some piecemeal examination, but there can be no assurance of finding them, short of a complete survey of all the preferential relations that obtain among the elements of a preference set, a survey that is beyond our capacities. Nor, short of a complete survey, can there be any assurance that a contemplated rectification of some intransitivity does not generate other intransitivities.

2.33 Maximizing and Satisficing

Our preferences do not satisfy the inscription thesis and therefore do not

[99] Imposing completeness, if it can be done, does not appear subject to an analogous problem. That is, imposing a complete ordering among some subset of the elements of a preference set is not liable to introduce *in*completeness elsewhere. The rub, of course, is "if it can be done." It is not clear that there is any procedure available to a finite mind for imposing completeness upon a preference set, especially if what must be completely ordered includes risky or uncertain prospects.

completely order our options. Hence, it is not true in general that we can maximize with respect to our preference sets since maximizing is not well-defined with respect to incompletely ordered preferences. Any of us can find ourselves in situations in which there is no answer as to which of our options best serves our preferences.

This can be over-stated or misunderstood, however. It does not imply that maximization is never appropriate, but rather that maximization is not always appropriate. For maximization to be a general requirement upon rational choice, it must be possible to apply it to any decision problem that can be constructed from the elements of a person's preference set. Regardless of the options with which the agent is faced, it must be possible (in principle) to identify one of them as being at least weakly preferred to all others. Denying that maximization is, generally, a rational requirement is consistent with maintaining that, for *some* sets of options for choice, one of those options may be weakly or strictly preferred to all others.[100] And, when this is the case, it may be entirely appropriate to hold that the agent should maximize with respect to her options in the given decision problem.[101] Thus, it is not true that denying that maximization is a requirement for rational choice is liable to infect all ordinary decisions, e.g., about what

[100] There are two ways this may be so. First, the actual set of options may be fully ordered by the agent's preferences (though not all possible sets of options would be). Second, some subset within the actual set of options may not be preferentially ranked with respect to each other, but there may be some option strictly preferred to any member of the unranked subset. Two further possibilities, neither of which, arguably, should be assimilated to maximization, would obtain when either (a) there is some mutually unranked subset of options such that each member of the subset is strictly preferred to any option in the complementary subset of mutually ranked options, or (b) when there is some option which is weakly but not strictly preferred to every member of a subset of mutually unranked options and at least weakly preferred to any other option that is not a member of the subset of mutually unranked options.

[101] A complication is that an agent may have adopted some action-guiding principle as a result of a non-maximizing decision (when no maximizing decision was available) and that the principle dictates a non-maximizing choice in a new choice situation in which a maximizing choice is available – i.e., in which some option is at least weakly preferred to all others. On the assumptions that it can be reasonable to adopt such a principle and that it should, at least *ceteris paribus*, govern decisions to which it applies, then it may be that the rational thing to do is to select an option that would, but for the principle, be strictly dispreferred to some available alternative.

to have for dinner.[102] Over limited domains, there may often be a maximizing choice and nothing I have said should be taken to imply otherwise.[103]

The point remains, however, that maximizing cannot be appropriate to all choices because it is not always well-defined what a maximizing choice would be. Further, the larger the scope of a choice – that is, the greater the extent to which it has effects which can be expected to be substantial and lasting – the more likely it is that maximizing will not be apt. Choice of a career or of a mate provide good examples, for in each case other decisions will in turn depend upon, will be altered or modified, will even be made possible or impossible, in consequence of the earlier decision.

Part of the point is that the further effects cannot be foreseen in detail and may therefore impinge in unforeseeable ways upon matters made relevant by one's other preferences. But so far that is only a problem of uncertainty. There is an additional dimension due to the fact that one of the features of long-term plans is that their execution makes a significant difference to what the person is doing over the term of the plan and that *the person herself* is altered in the process. She engages in different activities, spends time with different associates, and acquires different preferences as an indirect result of executing the plan. Importantly, some preferences relevant to the choice to adopt and execute the plan may be preferences the person *does not have* when the plan is adopted. The uncertainty involved runs deep: not only is the agent uncertain what the future may bring, she is also uncertain how the unknown future will matter when the time

[102] Nor, more broadly, is it the case that denying that maximization is a rational requirement is equivalent to denying that there are any rational requirements or *desiderata* when maximization is not appropriate.

[103] There is a further question: How does a domain get limited? Domain-limitation may be the result of *non*-maximizing choices – for example, that only options for dinner are to be considered and, of those, only the members of some short list.

comes. The larger the scope of a choice, the larger is the set of preferences that may be relevant, and the set of relevant preferences (assuming that set is well-defined – which it may not be[104]) probably no more than intersects with the *complete* set of the chooser's preferences at the time of choice.[105]

If, then, maximizing cannot be applied to all choices – and, ironically, is least likely to apply where we would most like some clear-cut decision procedure – what can we do instead? The most popular, and also I think the most plausible, answer (apart from a dogged insistence – or presupposition – that we *can* somehow manage to maximize) is that we should lower our sights and settle for *satisficing*.[106] The core idea is that the agent should seek and select an option that is *good enough*, rather than one that maximizes. Its most natural application is to cases in which an agent is still searching for an acceptable option. (It would normally make little sense to select a worse member of a set of options known to be available simply because it is still good enough.[107]) Then a satisficer, rather than trying to determine what option is best in terms of all her preferences together, delimits some range within which a decision problem arises – such as what to have for

[104] Set aside, for the moment, any concerns about how to determine in practice the membership of the set of relevant preferences. Then suppose that each of a pair of options would have different effects upon the preferences of the chooser such that, if one option is selected, the chooser will come to prefer *A* to *B*, whereas, if the other is selected, she will come to prefer *B* to *A*. Does it make sense to say that one of those preferences, to the exclusion of the other, belongs in the complete set? Surely, both are in some sense relevant and both have the same claim to be included, but if both *are* included, the preference set will not be consistent.

[105] Will the chooser have preferences about the ways in which her preferences are subject to modification in consequence of some far-reaching choice (which preferences can then feed back to provide additional criteria or desiderata for the choice)? Quite possibly, but there is no more reason to expect that these preferences will completely order her options than that her other preferences will do so.

[106] I was introduced to the term by Nozick (1981, 300), who cites Simon's 1957 *Models of Man*. The idea has been much discussed, both by Simon and others. See, e.g., Schmidtz 1995, Simon 1996/1969, and, without using the term, his 1990/1983.

[107] In special cases, it might – for example, if there is neither a best member nor any tied for best in the set of available options. See Schmidtz 1995, 42-43. Also, see note 101.

dinner, whether to accept a job offer, whether to buy a house or keep searching – and then settles upon criteria such that, if they are satisfied, an option would, by her lights, count as *good enough*. The options are compared in light of the antecedently established criteria, and the first to qualify as good enough is selected.[108, 109]

Much can be done in the way of formal analysis of satisficing, but I shall leave that to others,[110] except for noting the interesting point that it appears that any rationale for satisficing must itself be a satisficing rationale. An argument cannot be mounted that satisficing is the best we can do (given uncertainty and incomplete preference orderings), for, apart from the fact that 'best' may have no determinate reference in the face of incompleteness, its success would be its failure. If there *were* a sound general argument that satisficing, in our circumstances, is the best we can do, that would assimilate satisficing to maximizing. Satisficing would then be what maximizing under those conditions amounted to. Satisficing can only be a genuine alternative if its rationale is something other than that it is the best procedure for selection among options.[111] And if

[108] If the criteria turn out to appear *too* easy to satisfy, they may be revised upward, or if too difficult, then downward. In either case, what is "too difficult" or "too easy" is itself at least implicitly a function of a satisficing judgment – that the effort and resources devoted to the search is or is not good enough. See Nozick 1981, 300.

[109] There are indeterminacies, intransitivities and practical dilemmas to which a satisficer is prey. Her choice in favor of one option and against others may be shaped by the order in which questions are asked and considerations brought to bear rather than by the relative merits of the options. If we could, we would like to avoid such difficulties. *In principle*, the maximizer escapes them, but even at its best, the escape amounts to less than may appear. For a maximizer, choosing in the face of risk or uncertainty, the maximizing choice may be to select the best member of a limited set of options, consisting of, say, *A* and *B*. It may still be true that, had he considered a third option, *C*, he would have ranked it above both *A* and *B*. Being a maximizer does not protect an agent against the possibility that actual decisions may depend upon the order in which options are presented or upon other extraneous factors, rather than upon the relative merits of the options. More importantly, the promised escape from practical dilemmas is only an illusion in any case unless we can (always) be maximizers – which we cannot.

[110] See Schmidtz 1995, Chapter 2 and especially 55-57.

[111] Schmidtz says that satisficing can only be of instrumental value "because to satisfice is to give up the possibility of a preferable outcome, and giving this up has to be explained in terms of the strategic reasons one has for giving it up." (1995, 45) Though he makes it clear that he thinks that the strategic

the rationale is not that it is good enough, or satisficing, what could it be?[112]

Much can also be done in the way of providing a rationale for satisficing by exhibiting problems with the maximizing model, but that I take to be sufficiently complete for present purposes. What I shall do in the next section is attend to a feature of satisficing that in its turn suggests what I think is the deepest problem with standard, maximizing decision theory.

2.34 Means and Ends

How does a satisficing agent guide her action? Within some domain of concern, she selects as an objective some state of affairs which she believes can be brought about or promoted through her action. She is guided by her judgment that the selected state of affairs is good enough, that it answers satisfactorily or well enough to her desires and preferences. In other words, she selects a goal and, then, barring alteration of the goal itself, guides subsequent action with respect to that domain by what she understands to be its suitability for the promotion of that goal rather than by its suitability for maximizing the satisfaction of her preferences in general.

Thus, there are two distinguishable stages in the deliberation by which a satisficer guides her action. First, there is goal-selection carried out in light of the agent's

reasons for (sometimes) satisficing are rooted in maximizing from a larger perspective, the conflict with my view is more apparent than real, since he admits (46) that there may often be no optimum from a global perspective: an agent may have to make a choice when nothing unequivocally favors one option over another.

[112] The various axiomatized methods for choice under uncertainty do not provide alternative, non-satisficing routes to the selection of satisficing because they are all ways of identifying some maximand which completely orders options. The satisficer does not have any general procedure for inducing a complete ordering over options. (It is an interesting question for further exploration whether the selection of one of those methods for choice under uncertainty might presuppose satisficing in that there is no proof that one of those methods is best.)

preferences, but it is not assumed to be necessary either that selection of the particular goal or even that the selection of some goal or other (then and there) is a maximizing choice.[113] The fact that the goal is selected as being good enough, as answering well enough to her preferences (which will not normally fully order her options), has the important implication that it need not be abandoned instantly should something better or apparently better come along. Since it was not selected for being the best, even proof that it is not the best will not necessarily lead to its abandonment.[114]

Second, once a goal has been selected, action within the relevant domain is guided by its relation to that goal rather than by maximization. To take a simple example, an agent who has embarked upon an investment plan may have chosen to set aside a given percentage of her income every month. Having decided that, she does not reconsider what to do with that portion of her income, whenever an unanticipated opportunity for expenditure arises. She does not, in a typical case, ask whether she would really be better satisfied, all things considered, with new furniture.[115]

[113] Of course, it *may* be, but the satisficing agent has no general procedure for ordering her options so as to insure that a maximizing choice can be identified. See note 100.

[114] The deliberation relevant to abandoning a goal in favor of something else might be said to be *dissatisficing* in structure. It will be appropriate to abandon a goal when it turns out to be bad enough. For a satisficer, there will be a gap between barely finding something else to be better than a currently pursued goal and appropriately abandoning its pursuit.

There are interesting comparisons to be made with Joseph Raz's conception of authoritative reasons as pre-emptive: "the fact that an authority requires performance of an action is a reason for its performance which is not to be added to all other relevant reasons when assessing what to do, but should exclude and take the place of some of them." (Raz 1986, 46; emphasis in original omitted) Adopting a goal is analogous to recognizing an authoritative reason and pre-empts other reasons that would have been relevant had the goal not been adopted.

[115] Contrast what a maximizing agent is supposed to do. In the first place, it becomes less clear what it means to settle upon a goal. The maximizer may, of course, in light of all his preferences taken together, undertake to bring about some desired state of affairs, but whether this amounts to settling upon or having a goal is open to question. The problem is that, intuitively, in settling upon and then in having a goal, there is an element of inertia: the goal governs subsequent action but is not itself readily subject to reconsideration. If it is also granted that there may be some motivational state that falls short of constituting the having of a goal, then there is at least a potential gap between being in some way motivated by an

There is a clear sense in which the satisficer selects a goal and then guides her action by its relation to that goal. Whether or not the maximizer can be said in the same sense to have a goal or to guide his action in terms of a goal, the issue I wish to pursue is connected to whether he can distinguish his goals from the means appropriate to them – more precisely, whether he can distinguish the differing ways in which preferences with respect to outcomes and preferences with respect to the steps involved in bringing about those outcomes are relevant to his choices. If the distinction cannot be adequately drawn, then there will be a sense in which the maximizer *cannot* be said to guide his actions in terms of his goals.

Consider the following all-too-common problem. An agent has adopted a plan at a time, t_0, to bring about a preferred outcome at a later time, t_2. Execution of the plan requires performance of a particular action (the Step) at an intermediate time, t_1. I shall suppose that at t_1 there has been no change in relevant information available to the agent nor has there been any unforeseen change in the agent's preferences, but at (or just before) t_1, the agent strictly prefers not to take the necessary Step. In addition, we can suppose that the preference change with respect to the Step was itself foreseen when the plan was adopted.

Such situations are familiar. An example might be deciding upon a diet. There is an envisioned outcome, losing weight, ranked above other accessible future outcomes and

envisioned state of affairs and having it as a goal to bring about that state of affairs. We can expect the maximizer to be less attached to his goal than the satisficer, for he will be constantly ready to give it up should something better come along. Perhaps this degree of attachment is not sufficient for having a goal. On the other hand, it is of course true that even a non-maximizing agent is prepared at some point to reconsider, so readiness to reconsider alone cannot disqualify the maximizer as a goal-pursuer. The disqualification may be grounded in his being *too* ready to reconsider, but I do not know how to determine what degree of readiness is too great. Since none of the argument to come turns upon this being, by itself, an important difference, I shall not pursue it.

a necessary step, such as refraining from between-meal snacks. In addition, at the time the plan is adopted, it is recognized that there will be temptations to snack between meals: when the Step must be taken, the agent will prefer snacking to sticking to the diet. On one hand, it appears that the agent's reasons for taking the Step are just the same as for initially adopting the plan – no unanticipated information or preference has entered the picture. If the plan was initially well-conceived – that is, if it was reasonable to adopt it – the agent ought to take the Step. On the other hand, now that the prospect of snacking is immediate, the agent does *not* prefer the outcome expected from refraining from the snack. He would, right then, rather snack than lose weight. Why must he be bound by his preferences of a few hours earlier? If it is rational for him to guide his actions by his preferences, why are the preferences at t_0 decisive, while those at t_1 are discounted – especially since it is the preferences at t_1 that are actually experienced at the time the choice to snack or not must be made?[116]

Most of us – however difficult we find it to carry through in practice – suppose that the former argument is better: Having made a reasonable plan, and in the absence of relevant additional information not already taken into account in the formulation of that plan, it is reasonable for a person to take the necessary steps to implement the plan, even if those necessary steps are dispreferred at the time they must be taken.

However, according to standard decision theory, this misdescribes the situation. It is not that the first argument is invalid, but that it depends upon a false premise. In standard decision theory, the only reasons we have are based on preferences and expected

[116] Does he still *have* the preference to take the Step, even if it is not motivationally salient? Perhaps we should allow that he may, but if so there is at least an apparent conflict among his preferences, and it is not clear which should govern his choice.

consequences (subject to a budget constraint) at the time a choice is made. A decision cannot rationally depend – except insofar as this affects current preferences and expectations – upon a past event such as having adopted a plan. Thus, if the step needed to carry out the plan is such that one would prefer not to take it at the time of choice (i.e., when the step must be taken or not), then one has reason at that time not to take the step. But if this fact was really foreseen when the plan was adopted, the plan fails to be reasonable because its execution depends upon the taking of a step which it is not reasonable to take. One who accepts the rationality-defining postulates of standard decision theory should either not have formulated a plan aiming at that goal or else should have made provision that every step would be preferred to its alternatives at the time it would have to be taken. We can put this somewhat differently by saying that, for standard decision theory, reasonable plans are constrained by the requirement that they contain nothing but feasible steps, where feasible steps are all at least weakly preferred, when they must be taken, to their alternatives. If that requirement is not met, then the plan was not reasonable in the first place.[117]

This seems unsatisfactory. If standard decision theory is correct about situations of this sort, there may be an outcome which an agent would like to achieve and a plan that, if executed, would achieve that outcome, and it may be that if the plan were executed and the outcome achieved, the agent would be glad she had adopted the plan and taken all the necessary steps, but nonetheless, the agent cannot rationally adopt the plan because it incorporates infeasible steps. Her best available options are to either give up seeking that outcome or to undertake special arrangements to make sure that all the

[117] See McClennen 1990, especially chapters 12 and 13.

steps are feasible. Either option represents some cost, whether in the form of giving up the chance to obtain her most preferred outcome or in the form of making special provisions to avoid having to take infeasible steps.[118]

Why does this problem seem so difficult for the maximizer of standard decision theory? There are two points to note before trying to answer. First, the question is not (or not just) why *we* sometimes find it hard to carry through our plans. That seems adequately accounted for by imperfect rationality. Rather, the question is about why *ideally rational* agents would find themselves apparently having to settle for second-best. And second, it is specifically a problem for rational agents *as conceived by standard decision theory*. If, as I have been arguing, the conception of rationality embodied in standard decision theory is not normative for us, it may be possible to address the problems associated with taking steps to achieve a goal in ways not open to the maximizer.[119]

The reason the problem seems difficult, I think, is that standard decision theory has no satisfactory way of making a normative distinction between ends and means. If the distinction could be made, there would be conceptual room to hold that ends provide reasons for adjusting means but not *vice versa*. To see what the problem is, consider where or how such a normative distinction might be represented. There are two plausible candidates, that ends are to be characterized in terms of outcomes of actions or in terms of intrinsic preferences.

[118] A further concern is that the feasibility-insuring provisions might themselves be so costly that, if they are necessary to achieve the outcome, then the outcome is not worth achieving.

[119] Satisficers do not face the same problem, at least not in so acute a form, for they are not automatically subject to criticism for making non-maximizing choices and therefore not for taking counter-preferential steps. (I do not mean to suggest that being a satisficer is sufficient to deal with the problem in all its forms.)

Suppose we identify ends with outcomes and hence identify means with steps that contribute to bringing about those outcomes. Then, the problem is simple:[120] Though we can say what steps contribute to what outcomes, all normativity vanishes because the fact that a step contributes to an outcome will not provide any reason for taking that step or for avoiding alternatives. Any step and any combination of steps will lead to some outcome or other. What is needed, at minimum, is some way of discriminating *among* outcomes, to identify one or some as ends, rather than others, and therefore to enable the identification of some options, rather than others, as means to those ends.[121] In short, in addition to the identification of outcomes and contributory steps, something exogenous is needed to represent the normative force of ends.

It might be thought that the exogenous factor can be readily supplied. Consider *wholly derived preferences*, or *derivative preferences* for short. At a given street corner, I prefer turning left over turning right because I prefer one grocery store to another. If not for my preference between stores, I would (then and there) have no preference for turning one way over the other. On pain of infinite regress, however, not all preferences can be wholly derived; there must be some which are non-derivative or *intrinsic preferences*.[122] The proposal, then, would be that ends are to be characterized in terms of intrinsic

[120] More difficult problems pertain to questions about individuating outcomes and setting an appropriate time-horizon for the identification of outcomes, but those will not concern me here.

[121] And that is just the beginning, for many features of outcomes of action are not intuitively part of any end pursued in a given course of action. Typing rearranges small particles on my keyboard, but the arrangement or rearrangement is not what I aim at in typing. See also note 21.

[122] There may be single preferences, where *A* is preferred to *B*, which are wholly non-derivative in the sense that only the preference for *A* over *B* is relevant to any choice between the two. But it may be that a preference is not wholly derivative without being wholly non-derivative. The preference relation between the two may be part of some set of mutually supporting or interlocking preferences such that *A* would be preferred to *B* if nothing else were at stake, but that if something else were at stake, the preferential relations could be altered. Since no important part of my argument turns upon whether we are speaking about wholly or partially non-derivative preferences, I shall indifferently employ 'intrinsic preference' to cover both.

preferences. Ends will be intrinsically preferred to their alternatives, so, once ends are securely identified, we can turn to the consideration of contributory means – that is, we can show what derivative preferences an agent should have and act upon in light of his intrinsic preferences.

The problem with this is that no role is left for temptation. To return to the example of adhering to a diet, consider the (readily generalizable) case of George, whose end or goal is to lose weight. So far as he has only derivative preferences between steps or means, the only explanation for his not taking a step that is better than available alternatives at contributing to his ends must be in terms of misinformation, ignorance, or inadvertence. He will certainly have no motivation to take a step that either leads away from or less effectively toward his ends. But then, whence comes the temptation to snack? Surely, yielding to temptation is not a matter of an accidental misstep on the way to his goals.

The answer must be that George's preferences with respect to the steps to be taken are not wholly derived. He is motivated to snack rather than stick to the diet because he has some intrinsic preference for snacking, then and there. If so, there are two possibilities, that the preference for snacking either can or else cannot be integrated into a consistent ordering with George's other intrinsic preferences. If it cannot, then there is no consistent set of intrinsic preferences to identify as the relevant end or ends, and therefore none in terms of which to regiment means.

Matters are no better, however, if we suppose that George's preference for snacking can be integrated into a consistent ordering. For then, at least *prima facie*, the act in question is not, strictly speaking, one of *yielding* to temptation; rather, it is an act licensed by its service to his ends. There is an intrinsic, not merely a derived, preference for what we are calling "yielding" and, since ends are to be identified in terms of intrinsic preferences, no genuine yielding after all.

Now, it might be supposed that room for the possibility of yielding to temptation can be found in the thought that the snacking to which George is tempted is contrary to what he really or most prefers. Though his intrinsic preferences, including the preference for snacking, can be integrated into a consistent ordering, snacking then and there does not serve them.

This is very puzzling. Unless we wish to revert to a revealed preference theory, we should not complain that we cannot make sense of a situation in which, *ex hypothesi*, George can make a choice contrary to what he most prefers.[123] Nonetheless, there are difficulties, and though there are several possibilities, none seems adequate. To begin, what is the other element of George's preference for snacking: what is snacking preferred *to*? Presumably, it is preferred to not snacking. Also, however, sticking to the diet, which requires not snacking, is preferred to snacking. If that is not simply to amount to an inconsistent set of preferences and therefore to an inconsistent set of ends identified in terms of those preferences, there must be some sense in which the preference for sticking to the diet is what sets George's end while the preference for snacking does not. Since both the preference for sticking to the diet and the preference for snacking are intrinsic, we cannot distinguish between them on the basis of the presence or absence of an intrinsic preference. I have suggested we have to say that adhering to the diet is what

[123] Another possibility is that a revealed preference theory might be rejected on the grounds that it does not adequately accommodate indifference or incomplete preference orderings. However, in the case at hand, neither of these is supposed to be at stake. George is supposed to have a consistent preference ordering which is not served by snacking. One could object to the claim that *this* preference ordering (or one structurally like it) might not be revealed in choice behavior without being committed to the general claim that choice always reveals preference.

George most prefers, but how are we to understand that? We do not mean that the diet-adherence preference has greater introspectible intensity, for in those terms snacking may well be what George most prefers. Nor will it do to content ourselves, without further elaboration, with the formulation I gave earlier, that the act of snacking is contrary to what George most prefers, because, for the decision theorist, there are only formal limits on what preferences may enter into a utility function, and, subject to those constraints, any preference is to be considered on the same terms as any other. The set of his preferences is equally consistent if he snacks and alters the preference for adhering to the diet as if he refrains from snacking in order to adhere to the diet. What is needed is some further explication of the sense in which the preference for sticking to the diet is supposed to be of greater weight or importance than the preference for snacking.

That explication has not been, and, I submit, will not be, forthcoming. More precisely, it will not be forthcoming in terms that can be represented within standard decision theory, for the explanation being sought is one of the normative distinction between means and ends, not of the psychology or phenomenology of preference or desire. When we ask why adherence to the diet is more important than snacking, what we want to know is why George *should* abstain from snacking, and the answer to that lies in the fact that losing weight is George's goal or end. It is because losing weight is the goal that adherence to the diet is more important than snacking, not because adherence is more important that loss of weight is the goal. There will be no answer in terms of preferences alone.[124]

[124] Nor will there be an answer in terms of (just) preferences combined with beliefs and expectations. I do not mean, of course, that preferences (etc.) do not enter into the selection of goals, just that the role of goals or ends in the guidance of choice is not captured in those terms alone.

And that is the deepest problem with standard decision theory. Whether explicitly or not, the theory seeks to be reductive about ends, to account for them in terms of the satisfaction of preferences and the like.[125] But, as we have seen, the attempt to understand rational choice in terms of maximizing the satisfaction of preferences ultimately leaves the theory unable to express or represent the normative distinction between means and ends. Taking instrumental reasoning seriously requires that we go beyond decision theory.

2.4 Summary

Decision theory, understood as providing a normative account of rationality in action, given a set of beliefs, preferences and constraints, is often thought to be an adequate formalization of instrumental reasoning. As a model or representation of important features of instrumental reasoning, there is much to be said for it. However, if decision theory is to adequately account for or formalize (correct) instrumental reasoning, then its proposed axiomatic conditions must be normative for choice. That is, it must be that a choice is *rationally defective* unless it proceeds from a preference set that satisfies the axiomatic conditions.

Though some axiomatic conditions are largely uncontroversial, the same cannot be said for others. Accordingly, it is not easy to provide adequate support for the complete set of conditions. There seems to be no clear case that every agent who fails to

[125] What I envision in the way of a non-reductive account of ends is along the lines of Bratman's planning theory of intention. Though he does not typically speak in these terms, roughly, an objective or goal is what intentional action is guided toward, and an intention is "a distinctive attitude, not to be conflated with or reduced to ordinary desires and beliefs" (Bratman 1999, 10) – nor, I would add, should it be conflated with or reduced to preferences.

satisfy the complete set of conditions must be rationally mistaken. Indeed, the apparent fact that competent decision-makers, including experts in decision theory, make choices that would be disallowed by the axioms is one of the sources of doubt as to their normative standing.

For my purposes, the most important of the conditions is Completeness, the requirement that an agent's preferences completely order her options. Applied even to relatively small numbers of elements in a preference set, the number of comparisons required, if each must be made explicitly, quickly becomes too large to be plausibly managed. If extended to all the options that can be constructed from elements of her preference set under conditions of risk (to say nothing of uncertainty), a complete ordering would involve infinitely many pair-wise rankings. The agent cannot have explicitly performed all of these rankings, and so, if her preferences do completely order her options, it must be assumed that her preferences have an underlying structure which suffices to determine all the needed preferential relations. The claim that there is such an underlying structure to an agent's preferences, that a complete preferential ordering is inscribed in her preferences, is what I call *the inscription thesis*.

I focus upon Completeness and the related inscription thesis because, although we cannot satisfy the axiomatic conditions unless the inscription thesis is true of us, it can be shown to be either false or enormously unlikely that the inscription thesis is true of us. In neither case is it reasonable for us to believe it of ourselves. Further, if it is not true of us, there is little we can do to rectify matters. There are no obvious steps to take that would result in our coming to have a complete preference-ordering. The reasons that it is implausible to think the inscription thesis true of us are also reasons to think that it is implausible that we can ever bring it about that it will, in the future, be true of us.[126]

Since a complete ordering of all of our options is not inscribed in our preferences, maximizing cannot be a general rational requirement. Rational choice may include maximizing when one option is weakly preferred to all others, but it is not defined by maximizing, for it is not always well-defined what it is to maximize.

The most important alternative to maximizing is satisficing, in which an agent selects an option because it is satisfactory or good enough in terms of her preferences, the rationale for which must ultimately itself be satisficing in nature. Its importance is, first, that it provides an alternative to maximizing that is within our capacities, and second, that it provides a natural way to model the selection of goals or ends in light of an agent's preferences, without implying that the having, adoption or pursuit of goals is reducible to or explicable entirely in terms of the agent's preferences.

Having a non-reductive account of ends or goals is, in turn, important in order to have a satisfactory account of ordinary instrumental reasoning, including such commonplaces as the fact that we can be tempted to act in ways in conflict with our objectives. Though there is much that can be learned from decision theory, it does not adequately represent instrumental reasoning.

[126] Relatedly, even if there were steps we could take to impose Completeness on our preferences, it is not clear that we would have any reason to do so, for the supposed reason would either depend upon a complete ordering of our preferences or not. If the former, then the argument for imposition is fatally compromised, while, if it is the latter, the incompleteness of our preferences leaves open the possibility that the reason will be undefeated, untied, but still not rationally decisive.

CHAPTER THREE: THE SCOPE OF INSTRUMENTAL REASONING

3.0 Introduction

As argued in the last chapter, decision theory does not provide an adequate account of practical reasoning or even of instrumental practical reasoning. But instrumental reasoning itself deserves further attention on at least three counts.[1] First, it is very much a part of our ordinary experience that we adjust means to ends and regard various kinds of failures to do so, at least when the issue is clear-cut from the agent's standpoint, as irrational.[2] We say, for example, that people must regularly change motor oil if they want their automobiles to work properly or that they are mistaken to rely on the purchase of lottery tickets for their retirement plans. Instrumental reasoning has the

[1] I do not believe that the instrumental exhausts practical reasoning, but I shall not venture far beyond it for the two reasons that much more, and much more that is interesting, can be done with instrumental reasoning than is commonly thought and for the practical (and instrumental!) reason that limits must be set somewhere to the scope of the current project if I am to finish it.

[2] Perhaps it would be better to say that we regard failures to adjust means to end (again, from the standpoint of the agent) as subject to criticism on the count of their rationality. A charge of irrationality is the extreme case of such criticism, but there may be failures of rationality that would be better described as non-rational or as less rational than some alternative, rather than as irrational.

advantage of familiarity.

Second, instrumental reasoning seems non-committal, or perhaps better, minimalist, in what it requires us to assume or presuppose in the way of value theory or any necessary ontological underpinnings. Even those who are most skeptical of normative discourse find it acceptable to grade actions as being more or less appropriate means to given objectives and do not find that doing so requires the invocation of mysterious non-natural properties or cognitive powers.[3] In instrumental reasoning, we get 'oughts' that do not seem to have a problematic relation to what is the case.[4] Instrumental reasoning has the advantage of being metaphysically and epistemologically undemanding.

Third, as Nozick says (which may be partly explained by the second point),

> The notion of instrumental rationality is a powerful and natural one. Although broader descriptions of rationality have been offered, every such description that purports to be complete includes instrumental rationality within it. Instrumental rationality is within the intersection of all theories of rationality (and perhaps nothing else is). In this sense, instrumental rationality is the default theory, the theory that all discussants of rationality can take for granted, whatever else they think. (1993, 133)

Because it is the default position, the one that can be presumed to be shared, whatever else may be controversial, any results that can be reached in terms of instrumental reason

[3] See, for example, Mackie 1977, 27-30.

[4] I do not think this means that the conclusions we get are not genuinely normative or that their normativity is or must be reducible to something else. Failures of practical rationality are not to be identified with having made one or another sort of theoretical mistake.

should be acceptable to everyone. Instrumental reasoning has the advantage that it speaks to everyone.

Instrumental reasoning, then, has a number of attractive properties.[5] That, of course, is hardly enough to show that it is suited to play any large role in moral theory. To reach a judgment on that question – on whether and how instrumental reasoning is suited to play a large role in moral theory – part of what is needed is a more detailed examination of instrumental reasoning itself. What I shall do is to sketch the features that we ordinarily take to be involved in instrumental reasoning, beginning with the paradigmatic cases in which it is an external means to some single objective[6] that is under consideration. I will extend this with brief attention to cases in which more than one objective bears upon the selection of means, and then by considering how well the features elicited also characterize a less paradigmatic type of case that may be – and I shall argue should be – assimilated to instrumental reasoning. I shall then say something about the normative standing of instrumental reasoning in relation to the characterization I have given.

3.1 Features of Instrumental Reasoning

Instrumental reasoning is most clearly involved when the reasoning focuses upon

[5] A further feature that I find attractive is that, so far as the possession of the ends or objectives (in the light of which means are to be adjusted) is conceived in terms of motivating states of the agent, instrumental reasoning appears to satisfy the plausible internalist requirement that genuine reasons be motivating for rational agents. (For some doubts on this point, see Hampton 1998, Chapter 2.) Now, however, debates about internalism and externalism with respect to reasons have become increasingly intricate and the positions of both internalists and externalists ramified to the point that it is obscure what, if any, is the connection to the issues that originally prompted making the distinction (see, e.g., Darwall 1997 and Audi 1997). For present purposes, rather than become embroiled in that discussion, I set the issue aside.

[6] The terminology is explained below.

the relation of external means to some single objective. An objective, when considered in relation to an external means, is some action or state of affairs that can be specified independently of that means,[7] and that external means is something selected, adopted or performed because of its (expected) causal contribution to the realization of the objective. An objective may or may not itself be an end – something sought, aimed at or performed for its own sake – but that distinction will not concern me here, so I shall in the remainder of this chapter sometimes refer to objectives simply as *ends*, to external means simply as *means*, and to the relation between such an objective and the corresponding external means, when achieving the objective is the only relevant consideration, as the *simple instrumental paradigm*.

3.11 The Simple Instrumental Paradigm

To consider the simple instrumental paradigm, it is useful to employ a typology to distinguish the possible sorts of cases that satisfy its conditions. Four sorts are distinguishable, and they can be represented in this way:

	Sufficient	Not Sufficient
Necessary	Type I	Type II
Not Necessary	Type III	Type IV

The simplest imaginable case is one in which some means is causally[8] both necessary and sufficient to bring about or realize the end. In the second, the means is

[7] The objective need not be, and generally is not in fact, completely specified. There is normally a range of states of affairs that would count as the realization of the objective. The objective may be to have steak for dinner, but there are lots of different things that would count as having steak for dinner, and the agent normally will not have in mind such details as the exact size or cut of the steak or the precise microsecond at which dinner will commence.

[8] I shall not continue to qualify the means as *causally* (rather than, say, logically) necessary or sufficient for an end, but the qualification should be assumed.

necessary but not sufficient. In the third, the means is sufficient but not necessary. In the fourth, the means is neither necessary nor sufficient. Obviously, these logically exhaust the possibilities.[9], [10]

Cases of the first two types can be considered together because, so long as the end is not itself in question, but the means is necessary to realize it, the sufficiency or insufficiency of the means makes no difference to the points with which I am concerned here.[11] Consider a case in which there is some end, *E*, at which an agent aims, and some means, *M*, within his power, which is necessary to bring about *E*. The agent is aware of this and, being rational, selects *M* in preference to any alternative action, *N*, that may be possible in the circumstances.[12]

[9] In all of these, I am assuming, again, that no other considerations than the relation of the proposed means to their respective single objectives are relevant.

[10] For my purposes, the necessity or sufficiency of the means with regard to the end (or the lack of either) is to be assessed from the standpoint of the agent's knowledge or beliefs. Whether she is correct to have those beliefs is a further question from which I am abstracting. Additionally, since we are considering *causal* necessity or sufficiency, it will often not be the case that assignment of an example to one of the four classes is clear-cut, even abstracting from the agent's standpoint. In particular, it may be difficult to ascertain that some means really is necessary in the sense that no other means, including ones not considered, would work or that it is sufficient in the sense that, once the means is adopted, nothing could derail the expected realization of the end. Similarly, it may be difficult to ascertain whether some means that is assumed either not to be necessary or not to be sufficient really is not.

[11] There are two ways in which some means may be necessary but not sufficient to realize an end. First, it may be that a means, M_1, is not sufficient unless some other means, M_2, or some set of other means, $M_2, \ldots, M_n$, is also employed. Accordingly, when this is the case, there is some *set* of means in the agent's power which, taken together, is sufficient to realize the relevant objective, though no proper subset is sufficient. If that is so, this is just a more complicated form of Type I case; there is something the agent can do, namely, take all the means together, that is both necessary and sufficient to achieve the objective. Second and more interestingly, it may be that some means is necessary but not sufficient because the realization of the end depends in part upon factors beyond the agent's control, such as concurrent actions by others or the presence of other causal factors. These in turn may be factors about which the agent can do something, though, if the case is to remain distinguishable from the first possibility, and therefore in turn from cases of Type I, what the agent can do can at most raise the likelihood of the needed concurrence. (As Fred Miller pointed out to me, means that are necessary but not sufficient to bring about the realization of some end may not, by themselves, even increase the likelihood that the end will be realized. For example, if the end were to win a lottery, it would be necessary though not sufficient to go somewhere that lottery tickets are sold, but that does not by itself increase the likelihood of winning; one must also purchase a ticket.)

[12] At this point, there is a complication. In a particular situation, there may be no alternative to *M*

Consider now cases of Type III, where some means is sufficient but not necessary for the achievement of the end. There are two ways this might be so.[13] First, it may be that there is some non-zero probability that the end will be achieved, even if the means is not employed. However, there will also be some non-zero probability that the end will not be achieved if the means is not employed, since otherwise, there would be no sense to speaking of something as *means* to the end: to identify something as a means is to distinguish it as being in some way better than alternatives[14] in relation to an end. So, when there is some non-zero probability that the end will not be achieved without employing the means, the employment of the means raises the probability that the end will be realized – in this case, from some value less than unity to unity. If it did not increase the probability, the agent would have no reason (in terms of that end) to favor the employment of the means.[15]

open to the agent: *M* may be not only necessary to bring about *E*, but also the only thing the agent could do in any case, regardless of its bearing upon *E*. (This may be a case in which the action is over-determined. Had the agent not selected it – on his own, so to speak – some other factor would have intervened to result in his performing it anyhow. See Frankfurt 1969.) If so, then it would be unclear whether *M* was adopted as a means to *E* or not, since the agent would have performed *M* in those circumstances, even if *E* had not been his objective.

The problem is that the agent's deliberation and action are supposed to be rational, but "rational" takes its meaning in part from the contrasting case or cases in which deliberation or action is non-rational, irrational, or less rational. So, if the agent would have performed *M* whether or not *E* had been his objective – that is, if there is no contrasting case – in virtue of what is it or could it be true that *E* is his objective and that *M* is adopted *as* a means to *E*?

What is needed here is the truth of a counter-factual: If there were other options, alternatives to *M* that did not alter the type of case (so *M* is still necessary for *E*), then the agent would, if rational, select *M* rather than any of the alternatives. (And, by contraposition, if the agent would not or might not select *M*, then either *E* is not his end – at least not the only one that is relevant – or else he is not rational.)

[13] The two might also be combined, but there are no distinctive points to make about the combination.

[14] At least, something taken to be a means must be thought to be better than alternatives counterfactually available, as in note 12. In the present case, even that is not available for the relevant range of counterfactuals, for we are assuming that the end would be realized regardless of any selection of means.

[15] This does not imply that he has a reason *against* employing the means. He might be indifferent

Second, the agent might have available to him more than one option that is sufficient to realize the objective. There is available to him, say, M_1 which is sufficient to realize the objective, E, and M_2 which is also sufficient for E. Since M_1 and M_2 are each separately sufficient for E, neither is necessary. Here, there are two important points. One is that when there is some set of means, each of which is separately sufficient for the end, then, though none of them is necessary, the agent still has a reason in terms of the end for selecting one of them. It is not necessary that he select some particular one, but it is (rationally) necessary that he select one over any alternative that is not a means to the end. The other point is that when the agent is making a selection from among a set of separately sufficient means, *if* there is a reason for selecting one (or a member of some subset) over the others, then that reason must have other sources. There must, e.g., be some other end in terms of which one or more of the available means is judged less costly or more desirable than any other option among the set of available means.[16]

Much of what has just been said can, with appropriate modifications, be extended to cases of Type IV, in which available means are neither necessary nor sufficient to achieve the end. Again, it may be that the end could be realized whether or not some relevant means is employed, and again, there may be some set of available means from which a selection must be made. Still, this is perhaps the most interesting type of case covered by the simple instrumental paradigm because, in considering how it is rational to act in the selection of some means appropriate to the end, we encounter a feature pervasive in ordinary instrumental reasoning, the relevance of the probability that some

whether the means is employed or not.

[16] I do not mean to imply that nothing but some other end could be the basis for selection among a set of means.

selected means will result in achieving the given objective. Probabilities have, of course, figured in the earlier discussion of Type III cases, but then, their only role was to discriminate between some actions that do not involve employment of a means a given objective and others that do. Considerations of probability did not there bear upon selection *among* means. In none of the other three types of case did the sufficiency or insufficiency of the means for the end provide any basis for selecting among candidate means. It did not in the first and second, because there, the means was supposed to be necessary for the end, whether or not it was sufficient. It did not in the third because all of the candidate means were supposed to be sufficient.

What we have here is much more interesting. One way in which means can be neither necessary nor sufficient to achieve an end is when they stand in a probabilistic relation to the end. A means, M_1, may be more likely to bring about an end than some alternative, M_2. It appears that we can use the end, E, to judge that M_1 should be selected rather than M_2, even though neither of them is either necessary or sufficient for the realization of E.[17]

There is here an extraordinarily interesting question as to why, *for the single case*, it is reasonable to act on the basis of what is most likely to happen. We can say, of course, that the agent facing a long run, or an indefinitely extended series of cases of the same type, will do better if, in each of them, she predicates her decision upon what is most likely to happen (though there are complications even here). There are also cases in which the reasonable thing to do is to take both the more and the less probable outcomes

[17] There is a hybrid case intermediate between Types III and IV. There may be some means that is sufficient for the production of the end (as in Type III) and also some means not sufficient for the end but which makes it more likely that the end will be realized than if means are not employed (as in Type IV).

into account in deciding upon action – for example, by purchasing insurance against something less probable (and worse). But there are also cases in which such hedging of one's bets is not an option. Suppose the agent is offered, only once, a forced choice between a pair of gambles, G_1 and G_2, with the same (desired) prize associated with each. If she selects G_1, she will probably receive the prize; if she selects G_2, she will probably not receive the prize. We all believe she should select G_1, but why? She may not receive the prize if she does and may receive it if she does not. To say that the reason is that it is more likely that she will receive the prize by selecting G_1 is just to reiterate our belief that she should guide her actions by the probabilities. It does not really explain why she should. (Remember that she is not facing a long run or many cases of the same type.) It might be proposed that what she is really aiming at, her real objective, is the best chance of achieving or bringing about some state of affairs – namely, the one in which she receives the prize – but that seems to misdescribe the phenomenology. It is because she cares about the state of affairs, because getting the prize is her objective, that she cares about the probabilities bearing upon it; her concern with the probabilities is only derivative. I shall not pursue this further, but will take it for granted that we are correct to assume what we all *do* assume, that it is rational – indeed, rationally required rather than just that it is not irrational – to predicate one's actions upon probabilistically expected outcomes.[18]

Assuming that we are correct to regard a means, M_1, as better than another, M_2, when M_1 is more likely than M_2 to result in the realization of the objective, E, we can

[18] This issue first struck me about 1994. Peirce raised the issue in "The Doctrine of Chances" (1957/1878, 64ff.)

describe that by saying that the sufficiency of a means for an end is a matter of degree, with the degree given by the comparative probabilities, and that a rational agent will (*ceteris paribus*) select the more over the less sufficient means.[19]

3.12 The Normative Control Conditions

Two significant points emerge from the foregoing discussion of the cases covered by the simple instrumental paradigm. First, when a means, *M*, is necessary to the achievement of an end, *E*, since we have stipulatively denied the relevance of any other considerations, then the agent, if rational, must select *M* from among his various options. The end, *E*, serves as a principle of selection from among the agent's options (*M*, *N*, *O*, etc.) that, given the circumstances, uniquely picks out *M*. When a means, M_1, is not necessary to the realization of the objective, *E*, but is in some other way contributory to it, then the agent must, if rational, select either M_1 or some alternative, M_n, that does at least as well at contributing to the realization of the objective and must select one of these in preference to any option, *N*, that does not contribute or does not contribute as well to achieving the objective.[20] Options can be sorted, in terms of *E*, as better or worse (when

[19] It is plausible that if this is so, there is a significant further constraint on the correctness of instrumental reasoning. I have said nothing so far about the source of the probability judgments upon which an agent would have to rely, but whatever their source, sets of probabilities can fail to be coherent. Even without assigning definite numerical values, we can see, for example, that it is not possible that *A* is more likely than *B*, *B* more likely than *C*, and *C* more likely than *A*. So, if an agent should rely upon probability judgments in deciding what to do, he is less likely to achieve his objectives if the probability judgments upon which he relies are not coherent. Therefore, he has a reason, in terms of his objectives, for making his probability judgments coherent. This is of more than theoretical importance because people are not, in general, very good at assessing probabilistic reasoning. We not only make mistakes (which might be explained by carelessness, the difficulty of the assessment, or inadequate time), but *systematic* mistakes. For discussion of many of these, see Dawes 1988.

[20] This may require some qualification, in connection with issues about maximization and satisficing. In particular, it will be reasonable to adopt a means, M_m, which does not contribute as well to the objective, *E*, as some alternative, M_n, when it is not well-defined what is an optimum with respect to *E*. M_n, for example, may promise to make me wealthier than M_m, but it may be not be well-defined what an

M is necessary to realizing *E*, into one that is good as contrasted with all others, which are not).

Second, normative force flows from the end to the means, and not *vice versa*. There is reason for selecting *M* in terms of *E*, but none, either for or against *E*, in terms of *M*. Of course, *M* may be an objective which anchors further instrumental reasoning about what contributes to achieving or bringing it about, but that does not alter the point: *M* may have normative force relative to some further means, M', but the normative or reason-giving force still flows uni-directionally from end or objective to means.

Now, it might be thought that this is just an artifact of the stipulation imposed in the initial description, that there are no other relevant considerations. After all, in ordinary thought, we do consider the acceptability of ends in the light of means, do sometimes judge that an end is not worth having if it can only be had by means we are unwilling to adopt. A person may decline an opportunity to make money by cheating. On a larger scale, a research institution may accept stringent limits on experimentation on human subjects in the study of a disease, where such experimentation might credibly promise to advance the search for a cure.

This is all true, but I do not think it alters the basic point for two reasons. First, in at least some cases, putting matters in this way misdescribes what is going on. The objective may be, e.g., not just to make money, but to make money honestly. Isolating "making money" as the end to which various means, such as cheating, may be considered, may misrepresent the agent's actual end:[21]

optimum of wealth for me is, and therefore possibly not well-defined whether M_m or M_n would move me closer to the optimum.

[21] This has obvious connections to questions about constitutive means, which will be further

> Adopting something as a goal is not just a matter of attaching a positive value to its accomplishment and counting this in favor of any action that would promote it (unless this is overridden by considerations coming from elsewhere). When we "adopt a goal" we normally give that goal a particular status in our lives and in our practical thinking, such as the status of a long-term career objective, or of a whim, or of something that we want to do sometime on a vacation. That is to say, the intentions that constitute adopting the goal specify the kinds of occasions on which it is to be pursued, the ways it is to be pursued, and so on. So the limitations indicated by the qualification that other things must be equal include conditions determined by our understanding of the goal and the way it is a goal for us, not just limitations imposed by other values that might "override" it. (Scanlon 1998, 86)

If this is the sort of case under consideration, then it is not really an exception to the thesis that normative force flows uni-directionally from end to means. Only misdescription of the end makes it appear that it is being judged unacceptable in the light of the means. But, of course, there are other cases. An analysis of the above type cannot work when some restriction on how an end is to be pursued is not (once it is correctly described) part of the end itself. So, there is a second point: the stipulation that there are no other relevant considerations *is* doing substantive work. What it is doing, however, is serving to call attention to the fact that *if* some fact about or feature of the means makes a

addressed below.

difference to the acceptability or rationality of action in the service of the objective, the difference that it makes must have its sources elsewhere. That is, features of the means may make a difference, but not *just* insofar as it is a means; insofar as *M* is a means (and nothing but a means) to the realization of *E*, *E* provides or may provide a reason for (or against) the selection of *M* but not *vice versa*.

These two points, that an end serves as a principle of selection from among options available to the agent[22] and that reason-giving force flows uni-directionally from end to means, I shall label together as *the normative control of means by ends*. Such normative control, I believe, is quite generally characteristic of instrumental reasoning. Further, I suggest that the normative control of means by ends is not only characteristic, but both necessary and sufficient for an instance of practical reasoning to count as instrumental.[23]

This may appear truistic, but I know of no exceptions, neither of any clear cases of instrumental reasoning that, on one hand, fail to satisfy either of the conditions of normative control nor, on the other, of any further condition that seems essential for a tract of reasoning to count as instrumental.[24] If this is correct, then we can make use of the twin conditions of the normative control of means by ends as a marker to recognize instrumental reasoning in less familiar settings. In the next section, I shall briefly discuss what I take to be uncontroversial extensions of the simple instrumental paradigm.

[22] The options may, as noted above, only be counter-factually available if the agent has only a single option, that option being in fact a means to the objective.

[23] Such reasoning may, of course, have further features in virtue of which it also counts as some other form of practical reasoning. Some tract of deliberation may be more than instrumental without ceasing to be instrumental.

[24] On one level, I hope my characterization will appear truistic, for I do not intend to offer any further argument for it than to point to its presence in examples.

3.2 Extensions of the Simple Instrumental Paradigm

My main concern to this point has been with instrumental reasoning in which only a single end or objective bears upon action, and there has been only glancing or parenthetical reference, by way of *ceteris paribus* clauses or mention of other factors that would make a difference, were they present, to the possibility that more than one objective may be relevant to what it is reasonable to do. This has been, of course, deliberately artificial, aiming at eliciting the basic features of instrumental reasoning from the simplest available cases, but in much, perhaps most, reasoning that is uncontroversially instrumental, more than one end or objective is relevant.

3.21 Economizing

Perhaps the most familiar form this takes is what may be called *cost* or, perhaps better, *economizing*. I do not mean simply monetary cost, though that provides a useful example. When an agent has determined to pursue some objective, one question relevant to the selection of means is which will interfere least with other objectives. If he faces, say, a pair of options with regard to means that are equally good at promoting the objective, then if one of the options uses fewer resources than the other, resources which could otherwise be put to use in the service of some other objective, then the agent has reason to select that option.

Generally, the employment of some means to an objective requires that the agent give up something – time, effort or other resources – that could have been employed differently or on behalf of different objectives. What must be given up, which could have been employed differently, is the cost of pursuing that objective. Suppose that an agent has determined upon the pursuit of some objective, E_1, and that the only relevant alternative to E_1 – the only other objective that might be pursued with the same resources – is E_2. Suppose also that E_1 can be obtained or promoted equally well by two different employments of those resources, M_1 and M_2. Then, if M_1 and M_2 can be ranked against one another so that, after one of them is employed on behalf of E_1, what remains of the agent's resources is either more or else less satisfactory for the pursuit of E_2 than if the other had been employed, the agent has reason to select the employment of resources that is more satisfactory in terms of E_2. Though M_1 and M_2 are equally good in terms of E_1, one of them is better than the other in terms of E_2.[25]

To return from the abstract to the familiar, a person may have a set of options that are equally good at achieving some goal but one of which is cheaper than others. If so, she has reason to select a cheaper over a more expensive option. The reason this can work is that monetary cost is a useful proxy, for many cases, for what must be given up to achieve the goal. Had the money not been spent in achieving that goal, it would have been available for others. Or, had less been spent in achieving that goal, more would have been available for others.[26]

[25] A variation upon this pattern may occur in the time-ordering of pursuits. Of a pair of objectives, E_1 and E_2, a person might be better placed to achieve or promote both if she pursues E_1 first and E_2 second. Some resource needed for the pursuit of E_1 might become unavailable if E_2 is pursued first, but not *vice versa*. Another possibility is that the pursuit of E_1 before E_2 is better (or worse) in terms of some other objective, E_3. Time-ordering of pursuits is not, however, always best assimilated to economizing. For some cases, it may be better regarded as an instance of constitutive reasoning (to be discussed more fully below).

[26] Of course, not everything has a market price, and so, not everything can be compared to everything else in terms of relative market prices, but that does not affect the point that it is normally true that monetary assets could be employed differently in the service of some other objective. Goals that cannot be pursued or advanced through monetary means are, for that very reason, not in competition (along a monetary dimension) with goals that can be pursued through monetary means. The reason for selecting the less costly of otherwise equally good ways of pursuing a goal only depends on the assumption that there

3.22 Before Economizing

Though it is perhaps the most familiar, economizing or attention to costs is not the simplest kind of case in which more than one objective may bear on action. This is evident most plainly in the stipulation above that the agent has determined upon the pursuit of some objective. So far as the discussion has gone, it is only after some objective has been selected that cost enters the picture,[27] namely, the relative costs of different means for pursuing that objective in terms of other objectives. But questions can be raised on two fronts, first, as to whether there are ways that more than one objective may be relevant to a decision *prior* to the selection of some one of them for (immediate) pursuit, and second, as to whether and how, beyond that, the various objectives an agent has are relevant to the selection itself. Each deserves at least brief treatment.

3.221 The Relevance of Multiple Ends: Prior to Selection

Suppose an agent has some array of ends, E_1, E_2, and E_3, and that each of these represents one disjunct of a binary alternative: each will be achieved or else not achieved, and there are no intermediate degrees to which it may be promoted or advanced. Suppose further that this is a complete list: there are no other ends (nor any other considerations)

is *some* other goal that could be pursued through monetary means, not that all goals can be evaluated in terms of money.

[27] Cost may enter the picture in another way when it serves as a limiting factor on whether an objective is selected. For example, when I inquire as to the ticket price in order to determine whether to go to the concert, that presupposes that there is at least some other objective that bears on the decision. A case where cost is relevant to the initial selection of an objective is best uderstood in terms of the relevance of some combinatorial principle, as discussed in §3.222.

that bear on how she should act. Can we say anything about how it is rational for her to act in such a case without assuming that she has already selected one of them for immediate pursuit? Or, equivalently, since 'rationality' takes its meaning in part from the contrasting cases of that which is less or not rational, can we say anything about how it would be *unreasonable* for her to act? It depends on what options for action she has. Plainly, there are eight possible outcomes that could be correlated with options she faces. She can achieve

A. none of the ends, E_1, E_2, or E_3;
B. E_1 and neither E_2 nor E_3;
C. E_2 and neither E_1 nor E_3 ;
D. E_3 and neither E_1 nor E_2 ;
E. E_1 and E_2, but not E_3 ;
F. E_1 and E_3, but not E_2 ;
G. E_2 and E_3, but not E_1 ;
H. E_1, E_2, and E_3.

If we let the letters, *A* through *H*, stand for options she could have that would bring about the corresponding outcomes, then it would be unreasonable for her to select *A* if she has any other option and unreasonable for her to select anything but *H* if *H* is one of her options. In between, *E* and *F* are both better than *B*; *E* and *G* are both better than *C*; and *F* and *G* are both better than *D*. With no more information than has been given, there is no way to tell which, if any, is better or best of *B*, *C*, and *D* or of *E*, *F*, and *G*. Nor is there any way to determine how *B* compares with *G*, how *C* compares with *F*, or how *D* compares with *E*.[28]

Despite the indeterminacies just noted, the general point is that, given some set of

[28] There is an interesting parallel here to the economists' notion of Pareto-improvements. A change is Pareto-improving when it is advantageous to some – at least one – and disadvantageous to none. The same patterns of relative superiority between some options and indeterminacy between others appear.

objectives and some set of options that differently affect the realization of those objectives, there are or may be rational requirements pertaining to which of the options should be selected that do not presuppose that the agent has *already* selected some objective for immediate pursuit. There are ways in which an agent can be reasonable and also ways in which she can fail to be reasonable.

3.222 The Relevance of Multiple Ends: Combinatorial Principles

What, though, about the indeterminacies? This is a way of raising the question as to how to select one objective for immediate pursuit when the answer is not given simply by specifying the objectives and the agent's options. It seems that we often do make judgments of this sort, considering not just the fact that we have multiple ends which can be differentially advanced by the options available to us, but also considering their relative importance.

Suppose for the sake of illustration that the agent with the set of ends discussed above has only *B* and *C* as options. Assume also that the case differs in that the ends can be promoted to varying degrees. If she selects *B*, she will achieve one of her ends (to some degree); if *C*, then a different one (to some degree). Putting matters the other way around, selecting *B* is a decision against promoting E_2, and selecting *C* is a decision against promoting E_1.

If one of the options is to be rationally better than the other, then, in addition to the relevant ends, E_1 and E_2, there must be some combinatorial principle on the basis of which a choice of which to promote can be made.[29] This can take either of two principal

[29] I take no position here on whether the combinatorial principle is itself best conceived as being or

forms. It can be a priority rule according to which one of the objectives, such as E_1, is more important or valuable than the other and thus is to be promoted in the event of a conflict. Or it can be some weighting function that assigns greater importance or value to promoting, say, E_2 to *this* degree than promoting E_1 to *that* degree.[30] If it is the latter, there will be no implication that promotion of E_2 is always to be ranked ahead of the promotion of E_1; it may be that there is some degree to which E_1 can be promoted that would, if it were an available option, be more important or valuable than the promotion, to the degree possible in the circumstances, of E_2.[31] Whatever the form it takes, I think all that is required – that is, all that is required for it to *be* a combinatorial principle, though there are surely additional requirements upon its being reasonable or acceptable – is that it reduce indeterminacy. It need not decide all issues in order to genuinely decide some.

To this point, I have pointed to the need for one or more combinatorial principles if certain indeterminacies are to be overcome and have implied that not all possible combinatorial principles are equally acceptable, but have said nothing about why some particular combinatorial principle should be judged to be correct or better than some alternative. And I do not think we are in a position to do this here. At least some imaginable, if not very plausible, combinatorial principles can be ruled out as clearly

being based upon some end.

[30] There are other possibilities. One would be a hybrid principle that applies, say, a priority rule below some threshold and a weighting function above it. Still others can be imagined.

[31] There may be ways in which a weighting function can be represented as a (very complicated) priority rule or in which a priority rule can be represented as a weighting function. So far as I know, nobody is interested in the reduction of weighting functions to priority rules. It looks like a lot of work for no theoretical payoff. The possibility of reducing priority rules to weighting functions is more interesting but is crucial only if it is supposed, as many decision theorists might suppose, that rational choice in cases for which a combinatorial principle would have to be invoked must somehow be based on a weighting function. I think that is unlikely, but am content for the present to leave it an open question, while relying simply on the intuitive difference between weighting functions and priority rules.

unacceptable, such as the principle that tells an agent to pick one of his objectives at random and rank achieving it below all others or one that instructs him to list options in alphabetical order and select the first. Thus, we can say that the agent would be mistaken to select one of these principles, but that hardly narrows the field. In part, this is because they are so clearly unacceptable that no one is tempted by them in the first place, but more importantly, there are many, perhaps infinitely many, possible combinatorial principles that are not obviously unacceptable. All that can be said at this point is that *if*, in addition to some set of objectives, an agent accepts some applicable combinatorial principle, we will be able to say more about what it is reasonable for her to do, in terms of those objectives together with that combinatorial principle, than could be said in terms of the set of objectives alone.

In this and in the preceding sections, I have discussed the simple instrumental paradigm, where only a single objective bears upon action, and extensions of it to accommodate the relevance of multiple objectives. In the next section, I consider the relation of a more controversial kind of practical reasoning to an instrumental framework, *constitutive reasoning.*

3.3 Constitutive Reasoning

Consider constitutive reasoning.[32] A constitutive means to some objective is one that at least partially constitutes the objective to which it contributes. The objective

[32] Since part of what will be addressed in the present section is whether what I am calling constitutive reasoning is really practical reasoning at all, the label may appear question-begging, and it might be thought that quotes – "constitutive reasoning" – are more appropriate. I prefer to avoid cluttering the text with such devices in favor of noting the point here. My use of the phrase without the quotes may be taken as a promissory note to be redeemed by the subsequent argument.

cannot be adequately specified entirely independently of the constitutive means. A constitutive means can be contrasted with an external means in that, for an external means, its contribution to promoting or achieving some objective is causal and the objective can be specified independently of the means. The distinction will receive further discussion in the next chapter, and, for the most part, the points I wish to make here are obvious and, I hope, uncontentious. Still, in considering the simple instrumental paradigm and its extensions, only external means to objectives were in view. When constitutive reasoning is brought into the picture, some issues take on a new form or have to be reconsidered from a different perspective. There are two principal questions which tend to flow together. The first is whether constitutive reasoning is genuinely *reasoning*. The second is whether it is genuinely *instrumental*. Obviously, if it is not reasoning at all,[33] then it is also not instrumental reasoning. On the other hand, since it is not clear what it would be if it were not instrumental, doubts about its being instrumental are likely to spill over into questions as to whether it is reasoning at all. Of course, we can also run this argument in reverse to say that if it is instrumental reasoning, then it must be reasoning. Since I have already argued that the normative control of means by ends is definitive of instrumental reasoning, that is the path I will take: I intend to test the credentials of constitutive reasoning by examining whether it meets the normative control conditions. (Proceeding in this way has the advantage that I can set aside worries about how constitutive reasoning, if not instrumental, still qualifies as reasoning.)

[33] It is important for present purposes that it be *practical* reasoning, reasoning about what to do, and that is what I am assuming to be implied by the claim that constitutive reasoning is indeed reasoning. If constitutive reasoning were to turn out to be some kind of theoretical reasoning, that would be to no avail.

3.31 How Constitutive Reasoning Satisfies the Normative Control Conditions

The conditions of normative control, remember, were (a) that the end functions as a principle of selection from among candidate means, grading them as better or worse in terms of the end, and (b) that reason-giving force flows uni-directionally from the end to the means. Why might it be doubted that constitutive reasoning qualifies as instrumental by meeting these conditions? There seem to be three sources of worry here. The first is that there might not be constitutive *means* to an objective at all. The second is that an account that assimilates constitutive reasoning to instrumental reasoning runs the danger of trivializing the conception by construing instrumental reasoning so broadly that it would apply to *anything*. The third is that, if constitutive means partially or wholly constitute the objective to which they are means, it becomes unclear what either of the normative control conditions actually requires.[34]

The motivation for the first worry lies in the fact that, for anything to be an objective at all, something or other must be constitutive of it. There must at least be some state of affairs, believed or presupposed to be possible, which would be the realization of the objective. However, the fact that something or other would constitute the objective is not sufficient for part or all of what constitutes it to be a *means* to that objective. Means are adopted, selected or performed because of their (expected) contribution to achieving or promoting an objective. What reason do we have to suppose that something

[34] I have actually found only the second of the three raised in the literature. The remaining two are, so far as I know, my own (though Sarah Broadie [1987] touches upon the third, from a different angle and with different concerns than mine). I address all three in the attempt to pose the toughest challenge I can to the thesis – which I accept – that constitutive reasoning is best understood by assimilating it to the instrumental. (It is not difficult to find thinkers who *assume* either that there is no problem with the assimilation or else that no such option is available; it is much more difficult to find *arguments* for one or the other position.)

constitutive of an objective is really a means rather than just a constituent? Or, to put it differently, what is the import of the requirement that means be adopted *because of* their contribution to an objective? Why not say instead that there is just an objective, constituted by various actions, events or states of affairs, but that none of these are to be singled out and contrasted with others as means? On such a reading, the contribution of a means to an objective would then have to be understood as *causal* contribution, something that brings about or helps to bring about the objective, and so, all means would have to be external.

I think a full answer has to wait upon distinctions yet to be developed in response to the other two worries, but a beginning can be made here. The first point to note is that an objective may be some state of affairs to be brought about and with respect to which the only relevant actions are external means. There need be no *action* that is constitutive of the objective itself. (The objective is to have a flower garden; the planting, weeding and fertilizing may be external means.[35]) A means also may be a state of affairs, an objective, which is itself related to action only by way of further external means. (The lock on the jail cell is a means to prevent the escape of prisoners.) So long as this is the case, there is no place for introducing any notion of constitutive means.

Whatever a constitutive means is, it must be more intimately related to action than that. For a constitutive means to be distinct from a state of affairs to be promoted by external means (if at all), it must be an action, activity, practice or disposition with respect to action. Is this also enough? If, abbreviating, some action is part of the objective, is that sufficient for it to be a constitutive means to the objective? I think the

[35] They could be constitutive means if the objective were 'gardening' or 'being a gardener.'

answer is negative. Consider the objective of raising one's arm. There is an action without which one could not achieve that objective, namely, raising one's arm. This action is not some causal precondition for arm-raising; it just *is* the raising of one's arm. But it certainly sounds odd to say that one raises one's arm because of or for the sake of its contribution to arm-raising. Something more is needed if we are to make a respectable place for considerations of constitutive means to objectives, but to pursue it, we need to consider the other two worries.

Christine Korsgaard raises the second issue:

> But the instrumental principle is nowadays widely taken to extend to ways of realizing ends that are not in the technical sense 'means', for instance to what is sometimes called 'constitutive' reasoning. Say that my end is outdoor exercise; here is an opportunity to go hiking, which is outdoor exercise; therefore I have reason to take this opportunity, not strictly speaking as a means to my end, but as a way of realizing it. This is a helpful suggestion, but it should be handled with care. Taken to extremes, it makes it seem as if any case in which your action is guided by the application of a name or a concept to a particular is an instance of instrumental reasoning. Compare, for example: I need a hammer; *this* is a hammer; therefore I shall take *this*, not as a means to my end but as a way of realizing it. In this way the instrumental principle may be extended to cover *any* case of action that is self-conscious, in the sense that the agent is guided by a conception of what she is doing. (1997, 215-216)

An initial point to note is that what Korsgaard calls "in the technical sense, 'means'" is equivalent to what I have been calling 'external means.' This is only a terminological difference between us, turning on the fact that I see no reason to deny that some things that are not external means are nevertheless genuinely means. It is, however, a terminological difference that makes it more difficult for her to state the next point, for what she means by 'the instrumental principle' is a principle that affirms, roughly, that we have reason to take the means to our ends. If only external means are genuinely (or "in the technical sense") means, then it is awkward and possibly misleading to speak of extending the instrumental principle to cover constitutive reasoning. The extended principle would have to say that we have reason to take the means to our ends and also have reason to adopt ways of realizing our ends that are not means. (Why is that not just two principles rather than an extension of one?) If, on the other hand, we take a means to be an action taken or state of affairs selected or brought about for the sake of some objective,[36] then there is no problem with extending the instrumental principle to cover constitutive reasoning. It is extension only in the sense of recognizing a relevant similarity between cases that may be taken to be central – the employment of external means – and others that may be taken to be more peripheral – the employment of constitutive means – rather like the extension of the term 'mammal' to include cetaceans.

Setting aside terminological issues, Korsgaard still has an important point

[36] It is important that 'means to' and also 'ways of realizing' an objective be understood to be such from the agent's point of view; a means or way of realizing is selected or brought about because of the relation in which it is understood to stand to the objective.

pertaining to the danger of trivializing the conception of instrumental reasoning.[37] (It might be regarded as a version of the first worry viewed the other way around. The concern there was whether *anything* could count as constitutive reasoning. Here, the concern is, if we agree that anything counts as constitutive reasoning, whether we do not have to let in *everything*.) Her thought seems to be that what has to be avoided is the acceptance of some such principle as:

> (1) For any two things (actions, states of affairs, etc.), if one is an objective and the other a way of realizing that objective through action, then the second is a means to the first.

That principle, if accepted, would leave open the possibility that the two are identical. Then, if the objective happens to be the performance of some action, since the performance of that action is (of course) identical to itself, the performance will be a way of realizing the objective, and so, a means to realizing it. Every action undertaken in virtue of its falling under some concept or description will then be instrumentally rational. (And only the inadvertent or unintended will remain to provide lodging for lapses of instrumental rationality!)

Now, it is fairly plain what needs to be done to avoid this. Some restriction upon (1) is needed to rule out the possibility that the means and the objective are identical. So, it might be suggested that what we need is:

> (2) For any two things (actions, states of affairs, etc.), if one is an objective and the other a way of realizing that objective through

[37] Note that she does not put even this point as more than a reason for caution: describing ways of realizing ends as constitutive means to those ends can be taken to extremes; there is no suggestion that it has no place when the extremes are avoided.

> action, then the second is a means to the first, provided that the two are not identical.

I do not think that is quite adequate as it stands. At the least, it stands in need of clarification, but the best way to bring that out is to proceed to the third issue mentioned above. That was the problem of how to understand the conditions of normative control of means by ends when it is supposed that the means partially or wholly constitutes the end. How, if we cannot say what the end *is* apart from the means, can it be that the end serves as a principle of selection from among options available to the agent – that is, to distinguish means from non-means and better from worse means – and that reason-giving force flows uni-directionally from end to means?

To sort these issues out, something further needs to be said about the relation of constitution. First, we need to avoid misunderstanding what is being claimed when some means is said to be constitutive of an objective. A means is adopted for the sake of the objective, and, when the means is constitutive, it at least partially constitutes that objective. But we should not think of the relation of the constitutive means to the objective on *this* model: The objective is *A* and *B* together; the means is *A*.[38] To hold that some means is constitutive of an objective is *not* just to hold that it is a member of a set of elements, perhaps including other means and perhaps including some things that are not means, with the compound of elements being the objective in question. The basic reason is that the introduction of the compound end or objective would not do any theoretical work unless it made a difference and therefore was more than just a compound

[38] And *B* is either some other means or else something constitutive of the objective though not a means to it.

– if, e.g., it introduced some combinatorial function for relating its constituents.

Rather, there must be some way of understanding the objective *as* an objective, as something sought, aimed at or to be performed and in terms of which some criteria (which need not be complete) can be specified for whether something further – something else sought, aimed at or to be performed – counts as contributing to it, advancing it or being necessary to it, and therefore as a means to it. To use an earlier example, I may play tennis for the sake of exercise, and the tennis I play is constitutive of the exercise I get – perhaps, if I get no other exercise, wholly constitutive of it. At first glance, this looks as if it will violate the non-identity condition between objectives and means proposed in (2) above. The tennis I play and the exercise I get are identical: the very same events are, on the one hand, my playing tennis, and on the other, my getting exercise.

A closer look at the example will, I think, suggest the way in which (2) needs to be revised. Though the tennis I play is wholly constitutive of the exercise I get, I have some understanding of what exercise is independently of understanding what tennis is. Other activities than the playing of tennis (e.g., volleyball, walking, swimming or thumb-twiddling) can be considered as forms of exercise and can be compared with tennis along various dimensions – how strenuous they are, how appropriate to someone in my physical condition and so on. In terms of what exercise is, in combination with other relevant parameters, tennis can be assessed as better or worse exercise for me. The tennis I play and the exercise I get are *extensionally* equivalent, but *intensionally* distinct. In aiming to get exercise, I aim to do something that satisfies, at least reasonably well, the criteria by which I distinguish exercise from non-exercise and better from worse exercise. So, the principle needed to replace (2) is:

> (3) For any two things (actions, states of affairs, etc.), if one is an objective and the other a way of realizing that objective through action, then the second is a means to the first, provided that the two are not intensionally identical.[39]

Once it is recognized that the objective must be intensionally distinct from the means, even when the means partially or wholly constitutes the objective, there is no problem in seeing how the normative control conditions, that the objective serves as a principle of selection from among means and that reason-giving force flows uni-directionally from objective to means, can be satisfied by constitutive reasoning. Even when the objective is wholly constituted by the means, the two are only extensionally identical. The normative control conditions can be satisfied because the objective is not intensionally identical to the means.

3.4 In Search of Normative Underpinnings

We all find instrumental reasoning compelling. Whether it is a matter of causal contributions to or constitutive ways of realizing ends or objectives, we think that it is, at least *ceteris paribus*, rational to act on conclusions reached through instrumental reasoning and that we are subject to criticism on the count of rationality – that we are irrational or less rational – to the extent that we fail to guide ourselves by such conclusions. But why? What exactly is wrong with instrumentally irrational action? I

[39] The relevant descriptions in terms of which two items are judged to be intensionally distinct (or not) must be available to the agent in her acting and decision-making. See note 36.

shall try to provide a partial answer here, but, in approaching it, we need to see why there are problems (not, I think, insuperable problems) with the most obvious suggestion.

The natural response to questions about the normative force of instrumental reasoning is that instrumental reasoning has just the normative force of the ends from which it proceeds: we are justified in acting in instrumentally rational ways because otherwise, we will fail to realize (or are less likely to succeed in realizing) our objectives.

That natural response, however, may not be available to us. For consider the instructive analogy between correctness in instrumental reasoning and validity in logic. In a valid argument, truth is transmitted from premises to conclusion: if the premises are true, then the conclusion must also be true. Similarly, in correct instrumental reasoning, normative force is transmitted from ends or objectives to means. Instrumental correctness (call it *instrumental validity*) is a practical analogue of deductive validity.

Whether an argument is valid, however, is not the only, or sometimes the most important, question that can be raised about it. We can also ask whether it is sound, whether, in addition to being valid, its premises are true, so that we can be assured that its conclusion is true. One way to bring out this point with regard to logical validity is to add that falsehood is transmitted *backward* from the conclusion to the premises: if the conclusion is false, then so must be at least one of the premises from which it was validly inferred.[40] In a valid deductive inference, the truth of the conclusion follows from the assumed truth of the premises: the conclusion must be true, *given* the premises. But,

[40] This way of characterizing logic comes from Popper: "[D]eduction … is valid because it adopts, and incorporates, the rules by which truth is transmitted from (logically stronger) premises to (logically weaker) conclusions, and by which falsity is re-transmitted from conclusions to premises." (1965, 64) In this form, it is not sufficient to cover all systems of logic, since it does not address systems with truth-values additional to 'true' and 'false' or that employ, say, probability metrics. However, it is not necessary for my purposes to enter into these complications.

though validity is a matter of the relation between premises and conclusions, the truth of the premises or of the conclusions themselves is not. At least sometimes, there are ways of checking whether the premises or conclusions are true that do not depend on their place in a particular argument.[41]

And here we may encounter a disanalogy between instrumental and deductive validity. What, if anything, is the instrumental analogue of falsehood (or of truth)? Can there be any defect – call it *normative deficiency*[42] to put a label on it – of the conclusion of an otherwise correct course of instrumental reasoning that infects the end or objective with which the reasoning began? On one ground it might be thought that there cannot, for if there were any such defect, it would undercut one of the conditions of the normative control of means by ends, that normative force flows uni-directionally from ends to means. In terms of the means, there would be reason for rejecting, altering or qualifying the end.

This, however, would be a misunderstanding. First, when the normative control conditions were introduced, the claim was not that there could be no reason, based on some feature of the means, that could be relevant to the acceptability of the end, but rather that it could have no such feature *just insofar as it is a means* – that is, just insofar as it is contributory to the end or objective. But that does not rule out the possibility that the

[41] Some other argument, perhaps unstated, may be involved in checking. (I do not think it must be. Checking does not always involve inference.) From the premises, "all cows are green" and "Bossie is a cow," it follows that Bossie is green. But we can examine Bossie to determine that she is not green. It might be claimed that we are thereby relying implicitly on some such argument as "if Bossie were green, she would look green (given current lighting, etc.), but she does not look green; therefore, she is not green." Even if this is so, it is sufficient to point out that this is a *different* argument and, therefore, that the truth of the first argument's conclusion is not simply relative to the premises from which it was inferred.

[42] To avoid confusion, this normative deficiency should be understood as relative to context. Whatever the feature in virtue of which some conclusion of a particular course of instrumental reasoning is judged defective, that feature may not affect all courses of instrumental reasoning in the same way. It may be relevant *as a defect* in some situations but not in others.

means could be normatively deficient by virtue of some other feature. Second, to pursue the analogy with deductive validity, a defect in the conclusion of a valid argument – namely, that it is false – implies that there is a defect also in at least one of the premises. If no premise were false, then the conclusion would not be false, either. So, it runs the risk of misunderstanding to speak of the falsehood of the conclusion being transmitted backward to the premises, as if there were nothing wrong with the premises apart from the conclusion being or being found to be false. It is a misleading metaphor to say that the falsehood discovered in the conclusion infects the premises; rather, it shows that the premises were already defective. Similarly, if some defect, some normative deficiency, can be found in the conclusion of an otherwise correct tract of instrumental reasoning, that shows that there was already a defect in the objective or objectives[43] from which the reasoning proceeded.

So, in principle, we can admit the possibility of normative deficiency in the means to which we are directed by a course of instrumental reasoning, and, if we find it there, we will have to admit some normative deficiency in the ends as well. But this may seem not to be any progress at all, for it appears that all we have reached are *conditional* claims: *if* we find a normative deficiency in the means, then there must be some normative deficiency in the ends. That tells us neither that there ever is any normative deficiency in means nor how we find that there is. Instrumental rationality appears to be, as Darwall says, a matter of principles of *relative* rationality,[44] rationality relative to ends,

[43] If some combinatorial principle is employed and if that principle is not itself to be conceived as an end (see note 29), then the defect might be there instead. I shall assume that if combinatorial principles are not to be conceived as ends, then their role is analogous to rules of inference, the non-observance of which undermines the claims of a course of reasoning to be considered valid.

[44] Darwall 1983, 15-16.

desires or preferences that are given prior to or apart from any particular course or instance of instrumental reasoning in which they figure. But then, as Darwall also says, principles of relative rationality are principles of the *transfer* of reasons. Whatever reason there is for the objective is transferred to the means. That seems to get us no closer to saying that there ever *is* any reason for an objective. And if there never is a reason for an objective, then, if instrumental reasoning has just the normative force of the ends from which it proceeds, it has none.

Matters, I think, are not so desperate. We can say something about the normative force of instrumental reasoning without prior assumptions about the reasons that there are or may be for ends.[45] We can see this by reconsidering the parallel with deductive validity. In introducing the parallel, I noted that the truth or falsity of some premise or conclusion of a valid argument is not just a matter of its relation to other elements of the argument. But we can still recognize that, if we reach a contradiction in the conclusion of a valid argument, *that* must be false,[46] and therefore so must be at least one of the premises from which it was inferred. For the special case in which a conclusion is contradictory, no further checking is needed to determine that some premise is false.

Is there any analogue to this in instrumental reasoning, any way to recognize normative deficiency in the conclusion or premises of an otherwise correct tract of reasoning? I think there is. Consider that in instrumental reasoning the premises will include both some objective or objectives to be achieved and claims about how the world

[45] I do not mean that those questions are unimportant; I shall say more about them later. But the normative force of instrumental reasoning does not depend, at least not entirely, upon the answers to those questions.

[46] I omit from consideration paraconsistent systems of logic in which ~(P & ~P) is not a theorem.

is, about what is causally or probabilistically related to what. Then, there are at least two ways that we can recognize defects in the premises or conclusions of some course of instrumental reasoning.

First, the objectives from which the reasoning proceeds might be in conflict. This might be thought unlikely, but I do not actually think it is, especially when the reasoning is complex and takes into account the bearing of many objectives upon action. The reason is that testing for consistency is a problem subject to combinatorial explosion. If there are two propositions, *A* and *B*, then one test suffices – whether *A* is consistent with *B*. Add a third, *C*, and four tests are necessary – whether *A* is consistent with *B*, whether *A* is consistent with *C*, whether *B* is consistent with *C*, and whether *A*, *B* and *C* are consistent together. Add a fourth, and eleven tests are needed (which I won't detail). Matters only get worse from there. Since the consistency of a set of ends can be modeled in terms of the truth of a set of propositions (E_1 is achievable, E_2 is achievable, …, E_n is achievable), the same problem applies. No one who has a multitude of ends can reasonably be certain that they are all consistent. For the simplest case, a person might have set himself both to achieve and to prevent the achieving of some objective. Action taken for the promotion of *E* will defeat his attempts to prevent *E* and *vice versa*. If there is this kind of conflict between or among objectives, nothing he can do will serve to realize his objectives (though he may be able to realize one or some among them). That would surely be sufficient to show that he was mistaken to think or assume that *none* of the objectives in conflict (nor their combination) was normatively deficient.

Second, it may be that the pursuit of some objective or system of objectives cannot, though there is nothing inconsistent in the statement of the objectives themselves, be expected to be successful, due to the factual and causal relations that are also among the premises. For example, it may be that, given the way the world is, including in particular the way that people distribute trust, the single-minded pursuit of money is likely to have a lower monetary pay-off than some alternatives. Single-mindedly pursuing money would not then be a good way of getting money or not as good a way as, say, devotion to a career. If that is so, it would show that, *as the world is,* there is something normatively deficient in the single-minded pursuit of money.

We can identify a feature that is common to both types of case. They are, in different ways, cases in which action in the service of ends is *self-defeating*. A person who guides himself by certain objectives or sets of objectives, under certain conditions, either cannot succeed in realizing those objectives or is less likely to succeed in doing so than if his objectives were different. And this is, though minimal, a conclusion that can be reached without presupposing anything substantive about which objectives there is or is not reason to pursue. Accordingly, I suggest that being self-defeating is the appropriate practical or instrumental analogue to the defect in theoretical reasoning that is displayed in reaching (and maintaining) contradictory conclusions.[47]

In other words, instrumental reasoning has normative force *of its own* that does not depend upon prior assumptions about what it is reasonable to do or pursue. The claim that it has only the reason-giving force of the objectives from which it proceeds, and therefore none unless those objectives are presupposed to have reason-giving force, is

[47] In focusing upon the way that failures of instrumental rationality are self-defeating, I do not mean to be reducing something normative to something non-normative. If someone claims not to see what reason there is against a self-defeating course of action, I have no further argument to offer. (At least not of the present kind – I might be interested in arranging a series of bets as a contribution to his education!)

mistaken.[48]

3.5 Summary

Both because it is pervasive in and because it is perhaps the best understood department of our practical reasoning, instrumental reason deserves special attention from the constructivist. Centrally, what instrumental reasoning concerns is the relation between objectives, goals or ends and means – in the simplest and most paradigmatic cases, the relation between single objectives and means, actions or states of affairs selected for what is taken to be their causal contribution to bringing about the relevant single objective. More specifically, means or alleged means are graded as more or less (or not at all) appropriate to the relevant objective.

In these simple cases, we find two features which I label together as the normative control of means by ends, that an end serves as a principle of selection from among options available to the agent and that reason-giving force flows uni-directionally from end to means. I argue that these features, taken together, are both necessary and sufficient for a tract of practical reasoning to count as instrumental, and thus that we can make use of these conditions as markers to identify instrumental reasoning in cases that are less central or paradigmatic.

With the normative control conditions in hand, we can extend the account of instrumental reasoning to partial coverage of cases in which multiple objectives bear upon the selection of means and, by way of supplying a place-holder to stand for the

[48] I do not wish to suggest that there is nothing else to be said about the normative force of instrumental reasoning, just that at least this much can be said without appeal to any presupposed reason-giving force attached to the objectives that anchor tracts of instrumental reasoning.

results of further investigation, to the consideration of the relevance of some kind of combinatorial principles to more fully cover the bearing of multiple objectives. More importantly, I show that we can also comprehend within the account of instrumental reasoning what has been called constitutive reasoning, where the means deliberated about are taken to partially or wholly constitute the objective to which they are means, rather than simply to causally contribute to its production.[49]

Finally, I consider more generally the normative force of instrumental reasoning with a view to addressing the concern that it has no normative force that is not derived from some normative weight or importance attached to the objectives from which it proceeds, and thus, if no such weight or importance is presupposed, that instrumental reasoning falls short of being, genuinely, *reasoning* about what to do. In response to this concern, I argue that, even without substantive evaluative presuppositions about the ends or objectives to which instrumental reasoning is anchored, we can see that there is normative force in such reasoning, because failures of instrumental rationality are or tend to be self-defeating. Instrumental reasoning has normative force that does not depend upon value attached to the ends it serves.

[49] This is important for my purposes because the constitutive relation, as will become evident in Chapter Four, is needed especially for the eudaemonist account of the virtues. If the constitutive relation could not be understood as instrumental, that would at least be a source of further complication for the current project.

CHAPTER FOUR: THE STRUCTURE OF EUDAEMONISM

4.0 Introduction

Much of philosophic theory construction consists of the building of intellectual bridges between premises and conclusions, but neither in philosophy nor in engineering is the best strategy always to begin on one side and proceed to the other. Construction may proceed more fruitfully, working from both ends. Were we confident that we had all the right premises in place and could unerringly develop their logical and evidential bearing upon any conclusions that might be of interest, there might be little point in looking ahead to see where we might end up. Realistically, however, we often begin with some sense of where our premises will lead and work back and forth between premises and conclusions, trying to identify plausible premises for the conclusions we think are correct, as well as adjusting the conclusions in the light of what plausible premises can be found to support.

In the last chapter, I developed an account of instrumental reasoning that I think is well-adapted to provide part of the support for a kind of eudaemonism. In the present chapter, I begin construction from the other side, by trying to explicate the general features of eudaemonist theories. In the next chapter, I will try to establish the linkage between the two.

Eudaemonist theories form a family, united by resemblances, rather than a natural kind for which we might hope to provide an illuminating set of necessary and sufficient conditions. What is common to all eudaemonist theories, the normative centrality of living well or having a good life, can, with only a little ingenuity, be construed to apply to virtually any other moral theory as well.[1] But adding conditions to rule out non-eudaemonist theories does not help, for the plausible candidate-conditions would also exclude some theories normally understood as eudaemonistic. For example, if it is held that eudaemonists agree that the goodness of a life comprehends more than just its moral goodness, the Stoics are a significant exception. Or if the proposed condition is that eudaemonists agree that the importance of the virtues is not just instrumental to living well, the Epicureans are a significant exception.[2]

I take it that the project of coming up with a set of necessary and sufficient conditions for eudaemonist theories is fruitless. It may be impossible in principle, and, even if not, the analytical gains to be reaped from its completion are almost certainly not proportional to the effort that would have to be devoted. What I shall try to do instead is

[1] Other theorists would just differently specify what it is to live well.

[2] Julia Annas (1993, 339ff.) offers useful discussion and suggests that, on balance, it is not clear that Epicurus assigned a purely instrumental role to the virtues. However, it is sufficient here to note two points. First, there are passages that can readily be interpreted as assigning a purely instrumental role to the virtues – indeed, that are difficult to understand in any other way. Annas quotes, *inter alia*, "It is because of pleasure that we choose even the virtues, not for their own sake, just as we choose medicine for the sake of health" (p. 339). Second and more directly to the point, nobody takes Epicurus's supposed endorsement of a purely instrumental account of the virtues as a reason not to classify him as a eudaemonist.

to develop an account of the structural features of eudaemonism at its best.[3]

4.1 Some Preliminaries

For eudaemonism, the central moral conception is *eudaemonia*, where that is understood as an inclusive ultimate end of living well or having a good life,[4] and the moral virtues are understood as constitutive means to the achievement of that end. Though itself short, almost every clause stands in need of further explication, first, to forestall misunderstandings and second, to clarify more positively what the position is and involves. I shall first address three possible misunderstandings and then proceed to more substantive characterization.[5]

4.11 'Eudaemonia' and 'Happiness'

The Greek eudaemonists, including Aristotle, the Stoics and the Epicureans, posited eudaemonia as the over-arching aim in terms of which a good life was to be structured. The term has most often been translated as 'happiness,' though other

[3] I take my cue largely from Aristotle, though not, I hope, to the disregard of other eudaemonists, and when I seek quotes and arguments, illustrative of ancient eudaemonism, it will almost always be to Aristotle that I recur. Partly, this is due to familiarity and partly, is a matter of how I assess the relative importance of different versions of eudaemonism. Any who disagree with that assessment are free to take my discussion as outlining a possible structure for a eudaemonist theory. In any event, I shall hereafter employ the term, 'eudaemonism,' to denote the structure I delineate and will cease to qualify it as 'a version' or 'the best version' of the theory.

[4] Many eudaemonists and perfectionists have understood the ultimate end as something to be maximized – as living as well as possible or as having the best kind of life (see Hurka 1993, 55-57, for some references). For reasons discussed in Chapter Two, I think that is a mistake: maximizing is not a reasonable requirement.

[5] Since I am concerned to avoid certain misunderstandings of what eudaemonism is or must be, I shall in the next three sections frequently cite the ancient eudaemonists. This is for illustrative purposes, not because I suppose that all of them (which would hardly be possible) or any one in particular is entirely correct in moral theory. Rather, the citations serve as evidence that the misunderstandings I identify are indeed misunderstandings, since paradigmatic eudaemonists do not share them.

renderings, such as ‘flourishing,’ ‘well-being’[6] and ‘success’[7] have been favored by some.

On the whole, I think the traditional translation, ‘happiness,’ is unfortunate, though I would not go so far as to call it a mistranslation. The reasons can be brought out by attending to certain features of the way that we typically understand happiness, of how the ancients typically understood eudaemonia and the ways in which these contrast with each other.

The most basic of the contrasts – the remaining points amount to elaborations upon the theme – turns upon the fact that moderns often take happiness to be subjective, a matter of how one feels. Approximately since the time of Locke, it has been common to construe happiness as some function of pleasure (Locke,[8] Bentham,[9] Mill[10]) or as comprehensive satisfaction of inclinations (Kant[11]). Two near-corollaries are (1) that the question whether one is happy or not is something about which one cannot be or is at least most unlikely to be mistaken,[12] and (2) that happiness is or may be a relatively transient state – one may be happy *briefly*.[13]

[6] See, e.g., Nussbaum 1986, 6, and Richard Kraut’s comments in Aristotle 1997, 52.

[7] Austin 1970, 18-19.

[8] “*Happiness* then in its full extent is the utmost Pleasure we are capable of, and *Misery* the utmost pain: And the lowest degree of what can be called *Happiness*, is so much ease from all Pain, and so much present Pleasure, as without which any one cannot be content.” Locke 1975, Book II, Ch. XXI, §42, 258.

[9] “By utility is meant that property in any object, whereby it tends to produce benefit, advantage, pleasure, good, or happiness (all this in the present case comes to the same thing) or (what comes again to the same thing) to prevent the happening of mischief, pain, evil, or unhappiness….” Bentham 1973, 18.

[10] “By happiness is intended pleasure and the absence of pain, by unhappiness, pain and the privation of pleasure.” Mill 1965, 281.

[11] Kant speaks of the “concept of the sum of satisfaction of all inclination under the name of happiness” and says that “all people have … the strongest and deepest inclination to happiness because it is just in this idea that all inclinations unite in one sum.” Kant 1998, 399.

[12] Since being happy is a matter of one’s own feelings, the only barriers that can stand in the way of knowing whether one is happy are the barriers, if any, to successful introspection.

[13] A third feature of happiness, insofar as it is conceived by moderns in terms of pleasure,

By way of contrast, eudaemonia, as the ancients conceived it, though it was not of course divorced from pleasure or enjoyment, was more objective. It would make sense in their terms to say of someone, "he thinks he is eudaemonic, although he is not." And saying this would not just signal that he was unsuccessful in introspection and had misidentified his psychological state – which would be the most likely interpretation of our saying, "he thinks he is happy, although he is not." Rather, a person could be mistaken about whether he is eudaemonic in circumstances in which there is no question whether he has misidentified his psychological state. He might be correct about it but still mistaken in thinking it to be or to be part of being eudaemonic.

One of the ways in which it is clear that, for the ancients, eudaemonia includes more than subjective states is that it may be affected by events occurring after one's death:

> … both evil and good are thought to exist for a dead man, as much as for one who is alive but not aware of them; e.g., honours and dishonours and the good and bad fortunes of children and in general of descendants…. [F]or though a man has lived blessedly up to old age and has had a death worthy of his life, many reverses may befall his descendants…. It would be odd … if the fortunes of his descendants did not for *some* time have

enjoyment or satisfaction, is that it is essentially a *passive* state, a matter of what happens to one. On the other hand, eudaemonia is conceived as an active state, as itself active or essentially involving activity. Aristotle, for example, argues in several places that conceptions of eudaemonia not involving activity are defective (e.g.., *Nicomachean Ethics* [hereafter *NE*] 1095b 31-1096a 1, 1099b 18-24), defines eudaemonia in terms of "activity of soul in conformity with excellence" (1095a 16), and regularly assumes "that acting well is identical to happiness [eudaemonia]" (*Politics* 1325a 22).

some effect on the eudaemonia of their ancestors.[14]

Additionally, the more objective meaning of 'eudaemonia' (as compared to 'happiness') is evident in the fact that the former applies to a whole life, or at least to a substantial portion of it: "we must add, 'in a complete life.' For one swallow does not make a summer, nor does one day; and so too one day, or a short time, does not make a man blessed and eudaemonic."[15] Whether a person is eudaemonic at a time depends in part on what happens at other times. Again, by way of contrast, it does not sound out of place to speak of happiness as transient, to say of someone that she was happy (or unhappy) yesterday, without carrying implications about her happiness today.

Now, it would be going too far to say that 'happiness,' as moderns use the term, *cannot* be construed more objectively and therefore more in line with the ancient usage of 'eudaemonia.' The materials for an objective conception of happiness are as available to us as to the ancients.[16] With the necessary explanation, there need be no problem in referring to happiness in the explication of eudaemonism. My present point is only that, in order to understand what the classical eudaemonists meant, the additional explanation is indeed *needed*; otherwise, we risk distorting their meaning by importing and tacitly attributing to them more subjective conceptions which were foreign to their thought.[17]

[14] *NE* 1100a 18-31. I have substituted 'eudaemonia' for 'happiness.' Even if the particular example is not found compelling, it is still evidence that the ancients did not conceive eudaemonia as entirely subjective. Consider also the second clause, "as much as for one who is alive but not aware of them," and the further arguments it suggests about ways in which one's life may go worse apart from one's awareness that it is going worse.

[15] *NE* 1098a 16-19. I have substituted 'eudaemonic' for 'happy.'

[16] Charles Murray, for example, suggests "lasting and justified satisfaction with one's life as a whole." (1988, 44)

[17] Though I altered translations of Aristotle above to replace 'happiness' with 'eudaemonia' and 'happy' with the coined adjectival form, 'eudaemonic,' I will not continue to do so. I assume I have said

For my purposes, I think the spirit if not the letter of the meaning of 'eudaemonia' is well-captured by the phrase, 'comprehensively successful living.' A life is eudaemonic to the extent that it is successful in all the ways in which a life can reasonably be expected to be successful. It would, however, be cumbersome to introduce that phrase or variations upon it into any discussion of eudaemonia or the eudaemonic life. I shall normally just speak of living well or having a good life.[18]

4.12 Eudaemonism and Egoism

A second way in which eudaemonism risks being misunderstood may be encouraged by the first – that is, by the translation of 'eudaemonia' as 'happiness' without due attention to the fact that these terms must not be understood simply subjectively.[19] Since eudaemonism takes the conception of the agent living well as its central moral notion, it is sometimes suspected of or charged with being a version of egoism. But that, most philosophers think (and I agree), would disqualify it as a plausible moral theory.[20]

enough about the importance of understanding 'happiness' or 'eudaemonia' (if they are taken to be equivalent) objectively that there is no risk of distortion.

[18] Nussbaum suggests 'living a good life for a human being,' but often prefers to leave the Greek untranslated. (1986, 6)

[19] Even with a subjective construal of happiness, the inference from 'each person should guide herself by the pursuit of happiness' to egoism is not straightforward. Whether pursuit of one's own happiness is egoistic or not depends both on what counts as being in one's interests and upon what makes one happy.

[20] It might be urged that if the right content is assigned to interests, then, in the first place, eudaemonism can appropriately be classified as a form of egoism, because it will simply identify comprehensively successful living, including whatever that turns out to involve, with the agent's interests. In the second place, it may be urged that there is no objection to egoism so understood, at least not *because* it is egoistic; whether it is objectionable will depend (among other things) upon the actual content assigned to interests once the theory is fully worked out.

In the end, I do not think this response is adequate, for it trivializes the notion of interests upon which any theory deserving to be classified as a form of egoism must rely. Of course, if a theorist has a

> [A]ncient ethical theory begins with the agent's concern for her own life as a whole. Modern moral theories, by contrast, often begin by specifying morality as a concern for others; morality is often introduced as a point of view contrasting with egoism. If a basic and non-derivative concern for others is taken to be definitive of morality, then this contrast may be taken to show that ancient ethics is really a form of egoism; and this is indeed a frequent charge, and one that is often extended to modern versions of virtue ethics. (Annas 1993, 127)

There are at least two ways of expressing the concern here. One is a short argument, already suggested above, that eudaemonism must be essentially egoistic. The eudaemonist agent is concerned, first and foremost, as an ultimate end, with *her own life* going well. Anything she ought to do has to be explained in those terms; anything that it is wrong for her to do must in some way connect to its making *her life* go badly (or, at least, not as well as it otherwise might have).

That certainly sounds egoistic. Still, it is not sufficient to make the case. It fails to attend to the fact that, for eudaemonists, what it is to live well is an objective matter. And, to put it at its simplest, what it is objectively to live well may include not being an egoist. (And conversely, being an egoist may objectively make one's life worse.) Briefly,

completely free hand in specifying the content of interests, action in accordance with *any precept whatsoever* can be represented as in accordance with the agent's interests. But surely, more than that is wanted: it will not do to identify all moral theories as varieties of egoism.

This need not be adjudicated here, however. My concern is to argue that there is no reason a eudaemonist theory *must* be a form of egoism, not that it cannot be. It is not a response to that argument to defend the moral propriety of some version of egoism, since I did not claim to be showing that egoism is mistaken.

the egoist reads the concern with having a good life as "having a life that is good for the agent." But it may also be understood as "the agent having a life that is good;" that a good life is necessarily one that serves only or is directed only to the service of the agent's interests is not settled by any definitional arguments about the commitments of eudaemonism.

A more specific concern is that eudaemonism does not leave room for the right kind of concern for others; it does not, in Annas's words, leave room for "basic and non-derivative concern for others". Any concerns for others will have to be mediated through their role in facilitating a good life for the agent. Again, the key move in response appeals to an objective conception of a good life:

> There is no reason, *prima facie*, why the good of others cannot matter to me independently of my own interests, just because it is introduced as something required by my final good. The thought that is frequently suggested is that the good of others must matter to me just because it is the good of others, not because it forms part of my own good. However, there is no reason why this should be incompatible with its in fact forming part of my own good. For an ethics of virtue, the good of others matters to me because it is the good of others, and it is part of my own final good. It is quite unwarranted to think that the second thought must undermine the first. (Annas 1993, 127f.)

Here, I think Annas may slightly understate her case.[21] Her conclusion makes it sound as if she is only arguing that direct concern for others may be compatible with one's own good, which is certainly true. But a stronger thesis is open to the eudaemonist. It may be held that the agent can acquire a direct concern for the good of others *because* it is part of her own good to do so. The fact that she had a reason, other than the good of others, for acquiring a direct concern does not mean that she has not really acquired that direct concern. To *have* the concern is one thing; the reason for acquiring it is not necessarily the same thing.[22]

4.13 Centrality and Reductionism

There is a third way in which eudaemonism is liable to be misunderstood. I have spoken frequently of the normative centrality, for eudaemonism, of living well or having a good life. The term, *centrality*, was selected advisedly in contrast to saying that, for eudaemonism, living well is the *basic* or *fundamental* moral conception. The latter terms suggest, as the former need not, that all moral conceptions can be reduced to or explained entirely in terms of some prior notion of what it is to live well, a position that I call *reductionism*.

The contrast between centrality and reductionism is closely related to the well-

[21] I'm not sure whether she *does* understate it. That depends upon whether emphasis is placed on "required by my final good" or upon direct concern for others not being "incompatible with … my own good."

[22] Indeed, it hardly could be. If it is a concern one does not already have (if one *did* already have it, it would be inappropriate to speak of acquiring it), then, if one has any reason at all for acquiring it, that reason will have to be in terms of something other than the concern to be acquired. (Of course, it may be that one comes to have a concern without acquiring it for a reason.)

known contrast between conceiving eudaemonia as a dominant or as an inclusive end.[23] Conceived as a dominant ultimate end (corresponding to reductionism), eudaemonia is something single and independently specifiable to which all other objectives are subordinate, presumably because they are means to eudaemonia. Any other end has a role in a good life only by virtue of its service to eudaemonia and may – indeed, should – be abandoned if it ceases to be of service. However, conceived as an inclusive ultimate end (corresponding to centrality), eudaemonia includes other ends which at least in part constitute it. It is what it is in part because of the other ends it includes. Given that there are other ends which partially constitute it, abandoning one of them may *itself* spoil or detract from eudaemonia apart from any contribution that the abandoned end would make to something independently specifiable.[24] (To illustrate, if wearing a tie is in part constitutive of being well-dressed and so is wearing a belt, I can't make up for not wearing a belt by wearing a nicer tie. Being well-dressed isn't something independently specifiable apart from the various things that constitute it.)

The important point is not that one *couldn't* adopt a reductionist or dominant-end

[23] Hardie 1968, Ackrill 1980. See Broadie 1991, 198ff., for argument that it is not plausible to interpret Aristotle as holding a dominant end view. The distinction between dominant and inclusive ends focuses upon the relation, for the agent, between the ultimate end and action in its service. The distinction between reductionism and centralism is more theoretical, focusing upon the way that different ends or values are to be *understood* in their relations to one another.

[24] There is a point on which I shall not dwell here, but which, I think, weighs in favor of centralism. If there are multiple ends that partially constitute living well, it cannot be ruled out *a priori* that, in particular circumstances, there may be conflict among them and that, therefore, hard choices will have to be made that cannot or cannot unambiguously be said to be required or favored by considering what contributes to having a good life. On a dominant-end view of the human good, there can be psychologically hard choices (if it is difficult to bring oneself to act in the way one knows to be best) and epistemically hard choices (if it is difficult to figure out what is best). But *morally* hard choices, exemplified by Sartre's young man who must choose between joining the Free French and caring for his aging mother (Sartre 1975/1946, 354ff.), are not possible, for it will always be true either that one of the available options is best or that there is a tie for which is best. In morally hard choices, it appears that neither of (say) two available options (a) is better than the other, nor (b) are they equally good, nor (c) is there one which is not, for one reason or another, morally dubious. Such hard choices are neither as convenient for the theorist nor as comfortable for the agent, but seem to be a real feature of our moral experience.

model of the relation between eudaemonia and any other objectives, but that there is no necessity, simply because eudaemonism posits a single ultimate end as appropriate for the guidance of one's life, to do so. The ultimate end, instead of being dominant, may be *central* – it may be that in terms of which other ends and commitments are organized, integrated and understood. It need not, however, be the only thing that matters normatively, nor need it be the case that we cannot have some grasp of what matters normatively until we see those things in the light of their contribution to a life well-lived.[25] (For example, people surely have some understanding of the importance of virtues such as honesty, generosity or courage before they are able to think about their lives as wholes or in terms of what a good life is.)

4.2 Thinking about What it is to Live Well

If we set aside subjectivist, egoist and reductionist conceptions of eudaemonia,[26] how should the eudaemonist approach be understood? I shall try to begin to answer this in three stages. The first stage will treat of understanding in the most general way what it is to live well or to have a good life. For the next stage, I will pursue a more detailed account of the ways that eudaemonists have conceived of and classified the various kinds of ends and means and their inter-relations. Third, I will try to show how the moral virtues are understood and what their place is in the overall structure of eudaemonist

[25] On the theme I've been discussing in this section, I've found much that is useful in Hurley 1989, Chapter 2. Unfortunately for purposes of casual comparison, what I call 'reductionism,' she calls 'centralism.'

[26] My point in insisting on this, again, is not that no eudaemonist theory could properly be classified under one or more of these headings, but that none of them is a necessary feature of a eudaemonist theory. If we do not see that they are not necessary, we may find ourselves unable to recognize and appreciate what eudaemonism at its best can be.

thought.

Where should we begin in thinking about ethics? For eudaemonists, we begin with thinking about our own lives and raising the question, What is it for a person (for me) to live well or to have a good life?[27]

> We all think in retrospect about actions we have done and feelings we have had. For me to think about my life as a whole requires something further – I have to step back to some extent from my immediate present and projects, and think about my past and future. How have I come to have the projects I now have, and the attitudes I now have to these projects, and to many other things and people? To think about my life as a whole is to ask how I have become the person I now am, how past plans, successes and failures have produced the person who now has the present projects and attitudes that I have. And it is also to think about the future. How do I see my present plans continuing? Am I happy to go on living much as I have done, or do I hope, and perhaps intend, to change my commitments and attitudes?
>
> Ancient ethics gets its grip on the individual at this point of reflection: am I satisfied with my life as a whole, with the way it has developed and promises to continue? For most of us are dissatisfied with both our achievement and our promise, and it is only the dissatisfied who

[27] "Now it is thought to be a mark of a man of practical wisdom to be able to deliberate well about what is good and expedient for himself, not in some particular respect, e.g. about what sorts of things conduce to health or strength, but about what sorts of things conduce to the good life in general." (*NE* 1140a 25-28)

> have the urge to live differently, and hence the need to find out what ways of living differently would be improvements. (Annas 1993, 28f.)

This may seem a minimal starting point for thinking about ethics, but, spare as it is, it can serve to make at least two important points[28] relevant to further thought and investigation within a eudaemonist framework.

The first point to note is that *eudaemonia*, the notion of living well or having a good life, is introduced as a *thin* conception.[29] It is a conception of what (if anything) will satisfactorily answer to concerns about living well. It is, at the beginning, no more than that – a place-holder for something more richly specified that will, insofar as the inquiry is successful, take its place.[30] This is why it can both be "a platitude" that eudaemonia is the end (*NE* 1097b 21-23) and also a matter for dispute what eudaemonia is: "Verbally there is very general agreement, for both the general run of men and people of superior refinement say that it is happiness, and identify living well and faring well with being

[28] A third point of interest has to do with the relation between moral cognition and moral motivation. A person for whom a moral question arises may be seeking an answer to either of two different questions. She may, on the one hand, want to know how to tell what the right thing to do is, either in general or for a particular case. On the other, her question may pertain not to figuring out what is morally right, but to what reason she has, if any, to do what has been determined in some way to be morally right. On the plausible internalist view that what an agent has reason to do must be connected somehow to what she is motivated to do (see especially Williams 1990 and Korsgaard 1996a), the second question is in part asking what motivation the agent has to do what is right.

If the approach sketched above is an appropriate way to begin thinking about ethics, then the framework provides at least the beginnings of an attractive response to the motivational question. The motivation to do what is right, at least *some* such motivation, is implicit in the starting point since ethical reflection begins from and is itself motivated by dissatisfaction with how one's life is going (or at least by a concern to find out if one's life could be going better).

[29] "… [We] must presumably first sketch it roughly, and then later fill in the details." *NE*, 1098a 20-21.

[30] And understanding "eudaemonia" as a place-holder for something to be more richly specified does not even rule out the possibility that the inquiry will be *un*successful – that there will not turn out to be anything that satisfactorily answers to the relevant concerns. It may be that, even if improvement is possible on one or more dimensions, any improvement along a given dimension will be matched by non-comparable losses elsewhere.

happy; but with regard to what happiness is they differ…." (1095a 17-20)

One cannot, however, guide one's actions or life by a thin conception. Saying that one will do whatever it takes to live well is empty unless one also has some idea what it *does* take. In other words, what is needed is a *thick* conception of eudaemonia. This can be approached in two ways. In the first place, we can inquire as to what theoretical constraints there are on the notion. We can ask in effect: If there is anything suitable for specifying what eudaemonia is, what would it have to be like?

And it appears that we can make some progress along these lines. For example, whatever it is to live well, it would have to be sought, valued or desired for its own sake. If it were not, then either living well would not be sought, valued or desired at all (which is false by virtue of the starting point from which the question arises) or else it would be, whether directly or not, for the sake of something else that was in turn sought, valued or desired for its own sake. Then, questions would arise as to how to achieve that something else, whether what we *call* 'living well' is necessary to or supportive of achieving it, and why achieving that something else instead is not a *better* candidate to be identified with living well. Similar arguments can be adduced to show that living well must not only be valued for its own sake but must also be, in relevant senses, complete, final and inclusive.[31]

This much, and perhaps more, can be done in the way of formally constraining the conception of eudaemonia. Formal constraints, however, are not enough in at least two

[31] For arguments on all these points, see *NE* 1097a 15 – 1097b 21. I think these arguments are, by and large, defensible (or can be made defensible), but I will not devote time here to their defense. First, to do so would require considerable discussion peripheral to my present concerns, and second, it will be more evident just how a defense might proceed once we have in hand the distinctions related to the classification and inter-relation of ends to be introduced in the next section.

respects. First, they only specify conditions that would have to be met for something to count as eudaemonia or living well. They provide no assurance that anything *does* satisfy the conditions.[32] Second, at best, they operate as a filter for selecting among candidates that are generated in some other way. How might candidates for what it is to live well be generated?[33]

Here, we can derive some further lessons from the eudaemonist story about the starting point of ethical reflection. It is, in the first place, a story of a person who finds her life, considered as a whole, either unsatisfactory or doubtfully satisfactory. Anything presented as a candidate *to her* will have to promise to be more satisfactory. Eudaemonia, then, has to motivationally engage the beginner as a beginner. That means, of course, that somehow it has to make contact with motivations she actually has.[34] It can't tell her to systematically frustrate everything she wants or cares about – on the grounds that such frustration is what a good life 'really' amounts to. If it does, she will rightly ask, "If *that* is what a good life is, why should I care about having one?"

The initial motivational engagement begins with the agent's dissatisfaction with

[32] It is an interesting question what the appropriate response would be if one candidate satisfied all but one of the formal conditions we thought it reasonable to impose, while no other did as well. Should we keep looking for a better candidate, give up on the idea that anything satisfies all the conditions (and therefore that anything counts as eudaemonia), rethink whether the unsatisfied condition is really necessary or reinterpret the condition so it can be satisfied after all? More generally, when is something that is, though imperfect, the best realization of or the best approximation to a realization of a concept, *good enough* to count as a realization of that concept?

[33] Part of the answer will appeal to what the agent already believes about how a person ought to live, about what kinds of actions, aims and character traits are good, right, admirable or acceptable. This is not a way of saying that these beliefs are sacrosanct or beyond criticism or revision, but it is a fact that the agent addressed by a eudaemonist theory typically already possesses such moral beliefs, and, in ethics as elsewhere, we must start where we are. This is an important theme, but not one I shall treat here. I discuss it somewhat further in talking about the socially embedded character of the virtues in Chapter Five.

[34] The motivations she already has, of course, need not be entirely self-interested. There are reasonably well-understood evolutionary reasons that our original motivational complement includes some measure of direct concern for the well-being of others, especially of close kin. Additional other-directed concerns are generally inculcated and reinforced in the processes of maturation and socialization.

(or doubts about the satisfactoriness of) her life. However, it need not be supposed that her dissatisfaction is all there is to the motivational constraint, as if the agent were saying only: "I want something different; *this* is something different; so I'll pursue it." The motivational pre-conditions may be relevant to her selection in at least two further ways. First, she may find some particular proposal for a conception of eudaemonia implausible, to amount to saying that she must, to achieve it, systematically frustrate what she wants or cares about. For instance, she may think that Epicureanism, in affirming that only pleasure is good and that only pleasure and what is conducive to pleasure is desired in the ideal eudaemonic life, demands too drastic a change in what she already thinks worthwhile. Or to take another instance, she may say, with Aristotle, that any theory, such as the Stoic, which holds that a truly virtuous person can be eudaemonic even under torture, is something that could only be held by someone "maintaining a thesis at all costs." (*NE* 1095b 31-1096a 2)[35] Second, as will be discussed somewhat more fully below, eudaemonist theories must provide some developmental account of the process by which one comes to embody a conception of eudaemonia, and the agent may find the developmental account motivationally implausible – which is to say that she cannot see how *she* could follow through on the prescribed steps alleged to lead to living well. Implausibilities of these types may or may not be decisive, for they might be overcome by further considerations, but that does not imply that they are not real motivational constraints bearing upon the acceptability of particular eudaemonist theories. Where

[35] Aristotle, of course, is not thinking about the Stoics, who came later. Presumably, he has in mind the Socratic doctrine that no genuine harm can come to a good man (and therefore that anything that can happen to a good man, such as torture, cannot be a genuine harm).

relevant, they *need* to be overcome.[36]

Now, it is not likely that there will be some simple, obvious and readily takable step that will lead immediately to the more satisfactory state sought by the agent. If there were, she would already have taken the step or at least would be preparing to take it and so would no longer be in search of a conception of living well. Additionally, though I have spoken of the state or conditions she is seeking as satisfactory or more satisfactory, there is no reason for her to suppose that it will satisfy her *as she is now*. As she is now, she is not satisfied, and none of her options lead immediately to being satisfied with her condition. She must expect that the content of her satisfactions – that is, what would satisfy her if it were realized – will itself change if and to the extent that she finds and embodies in her life an adequate conception of eudaemonia.

These points suggest three further requirements. First, there must be some developmental or transformative process[37] that will lead the agent from where she is now to the kinds of motivations she would have if she were in fact living well.[38] Second, the content of her satisfactions in that envisioned condition must be such that achieving or having whatever would then satisfy her appears feasible. There would be little point in deliberately undergoing a developmental process that would alter one's motivations and responses so that, with the altered motivations, one could not be satisfied with what one

[36] In addition, the acceptability of a eudaemonist theory may be constrained in other ways than motivationally. As mentioned above, there are formal constraints, and there are also constraints derived from prior belief, especially widely shared prior belief. See *NE* 1145 b1ff. and note 33.

[37] For discussions of the developmental processes envisioned by ancient eudaemonist theories (and of much else), see Nussbaum 1994.

[38] For this reason, I do not think it needs to be true that the seeker must, at the initial stage, clearly see (what is said to be) the realization of eudaemonia as attractive. On one hand, it must be attractive in some sense or to some degree – otherwise, it would appear only as a way of telling her that she must systematically frustrate what she cares about, but on the other, the motivations of the practically wise person may be, at least partially, opaque to the beginner.

could then reasonably expect to achieve or have. Third, given that, due to the accidents and risks of ordinary life, the developmental process may or may not be completed,[39] she must see the steps to be taken as worthwhile, either because of the attractiveness of the terminus of the developmental process (which is identified with eudaemonia or living well) or in terms of what she sees to be desirable at the stage at which the steps must be taken (or, of course, some combination).

The general lesson to which these considerations point is two-fold. Any credible eudaemonist theory will have to meet certain motivational conditions and will have to propose, perhaps sketchily, a plausible developmental account, subject to whatever motivational requirements are relevant at any given stage, of the way in which one may move from the position of the beginner toward something that can be identified with living well. There is room for variation in the details and in the relative emphasis accorded to practice, habituation, imitation of others (presumed to be more advanced), participation in social and political life, and to reflection and discussion, but however the details are worked out, the (motivationally constrained) developmental story must be present. We must begin from where we are, cognitively, motivationally and affectively. Anything that we will be able to recognize as living well must be reachable or approachable by addressing what we care about and what concerns us as we are.[40]

[39] I abstract here from two further questions about the completion of the developmental process. The first has to do with ways in which the agent may be at fault for its non-completion and the second with whether the end-point is conceived as realistically achievable or instead as an ideal to be approached. With regard to the latter, note that if eudaemonia is conceived only as an approachable ideal, the question about the desirability of the intermediate steps becomes more urgent.

[40] I do not think this begs any questions against the possibility of purely rational motivation. If we are such that we can be motivated independently of particular desires, wants or preferences that we only happen to have, then that is a fact about "where we are" and therefore about what we have to work with in addressing our concerns or what we care about.

4.3 The Classification of Ends and Means

There are ways of acting, objectives that are sought, states of affairs that may be achieved, as we say, *for their own sakes*. We find these to be somehow worthwhile, desirable, satisfying, enjoyable, even without any appeal to something beyond them to which they are thought to contribute. These, I shall call *ends*. There are also ways of acting, objectives sought, states of affairs that may be achieved through action that are engaged in or sought for the sake of something else. These can be called *means*. For each person, it is in terms of the network of his ends and means, combined with what he believes about their relations to the world and to one another, plus any reasoning that may be brought to bear, that his actions are shaped.[41]

Indeed, unless our ordinary understanding of action and motivation in ourselves and others is radically mistaken, this must be the case. So long as we are reflective beings and so long as the ways in which we act are subject to reflective control, we can ask of any particular objective or action why we seek it, engage in it or practice it. To ask the question is to ask whether the objective or activity is an *end*, and so does not require any further justification or rationale, or whether it is a *means*, and so is warranted (or not) in terms of something to which it contributes. Broadly, there are three possible answers. We can appeal to something else as supplying our reason, in which case we can repeat the same question with respect to that. We can fail to find any reason, in which case, since what is being considered is something that (we are supposing) is subject to reflective

[41] Other kinds of action, such as the merely habitual, may be possible, but even the merely habitual, for normal adults at any rate, is under at least counterfactual control in terms of the agent's ends: it could be altered if some reason were recognized for the alteration.

control, we will cease to act in that way.[42] Or we can conclude that no further and distinct reason is needed – that the action, objective or practice is worthwhile for its own sake.

This may suggest an overly simple picture, though, in which means and ends are neatly dichotomous. It needs explication and complication in more than one direction. Here, I shall offer a classification of ends and means that reflects the richer structure explicit or implicit in eudaemonist theories. It is best to begin by presenting relevant distinctions quickly and then back-tracking to fill in details.[43]

- An *objective* is something sought, aimed at or to be performed.
- An *end* is an objective sought, aimed at or to be performed for its own sake.
- A *means* is an action taken or state of affairs selected or brought about for the sake of some objective.
- An *external means* is a means adopted for the sake of its expected causal contribution to an independently specifiable objective.
- A *constitutive means* is a means adopted because it is taken to at least partially constitute the objective to which it contributes. The objective cannot be adequately specified entirely independently of the constitutive means.

[42] We could, perhaps, fail to find a reason but also fail to find any reason for changing. I can see four interpretations: (1) It could mean that one is indifferent between a pair of options. But then we can repeat the question with respect to the disjunction of the pair: why do or aim at *either*? (2) It could mean that the two are thought to be non-comparable, in the sense that they are not equally good, but also that neither is better than the other. But this is to say that there are reasons for each that we either cannot rank or do not know how to rank against one another, rather than that there is no reason for the aim or activity at all. (3) It could mean that the two are equally good specifications of something else sought or aimed at. Again, this would not, except in a Pickwickian sense, be a case of there being no reason, but rather one in which the available reason underdetermines which is best. (4) It might be an expression of some kind of nihilism or skepticism which denies that there are any reasons for any action. That, I take it, is not a possibility admitted by our ordinary understanding of action and motivation.

[43] All of the terms distinguished here are to be understood as applying within the network of means and ends that characterizes the action, motivation and deliberation of a single agent.

- A *final end* is an end that is not sought or aimed at for the sake of any further objective.
- An *ultimate end* is a final end to which all other objectives are means.[44]

Using this terminology, three simple and fairly obvious points can be rehearsed quickly. First, an objective can be either a means or an end. An objective is just an object of intentional action which, in one direction, may or may not be sought or performed for its own sake, and in the other, may or may not anchor further deliberation with regard to the means suitable for realizing it. Second, there can be indefinitely lengthy chains or series of means linking action to some end that it serves. Third, being a means and being an end are not mutually exclusive. An objective may both be aimed at for its own sake and for the sake of something further to which it contributes.[45] There are further issues, however, that are not so quickly settled and that require additional consideration. These have to do

[44] Eudaemonists have often not distinguished sharply between final and ultimate ends, but ignoring the distinction leads to trouble. For example, in *NE* I.7, Aristotle says:

> Since there are evidently more than one end, and we choose some of these (e.g. wealth, flutes, and in general instruments) for the sake of something else, clearly not all ends are complete ends; but the chief good is evidently something complete. Therefore, if there is only one complete end, this will be what we are seeking, and if more than one, the most complete of these will be what we are seeking. Now we call that which is in itself worthy of pursuit more complete than that which is worthy of pursuit for the sake of something else, and that which is never desirable for the sake of something else more complete than the things that are desirable both in themselves and for the sake of that other thing, and therefore we call complete without qualification that which is always desirable in itself and never for the sake of something else. (1097a 25-35)

This seems to admit that more than one end might be complete without qualification (that is, that more than one might be final in the sense I indicated above), but if more than one end is sought only for its own sake and not for the sake of anything else, then the condition of self-sufficiency ("that which when isolated makes life desirable and lacking in nothing" 1097b 15-16), that Aristotle also thinks applies to the chief good, will not be met. There might be two final ends, each sought for its own sake and not for the sake of anything else. To nominate one of these to the exclusion of the other as the chief good would leave the life lacking in something, namely, in whatever is comprehended under the other final end.

[45] To hold that some objective, which is a means, is also an end is to be committed at least to the claim that it would *still* be aimed at, under some relevant counterfactual conditions, even if it did not contribute to something further.

with final ends, with ultimate ends and with the distinction between constitutive and external means.

4.31 Final Ends

We can begin by addressing the question whether, for a given agent, there must be any final ends. An argument was provided earlier that, given our ordinary understanding of our action and motivation, we must regard some objective (at least one) as an end. That does not settle the current question because an objective may be both an end and a means. So, it might be that all ends are also means and thus, that none are final.

There are two salient possibilities here: Action might be structured in the service of an infinite sequence of ends,[46] or there might be some finite cycle of ends, each of which contributes to and is contributed to by some other. The first can be ruled out for finite agents such as ourselves. Even if, in some sense, an infinite sequence of ends is possible, it is not possible for us: we would be unable to guide our actions in terms of such a sequence.

The more interesting possibility is that an agent might guide his action in terms of some finite cycle of ends. Perhaps he eats to work and works to eat. It is not altogether easy to be clear just what is supposed to be envisaged here. It cannot mean that the agent eats only in order to work and works only in order to eat, for then, eating and working would not be *ends*. Nor can it mean that eating and working jointly constitute what he

[46] Aristotle, at 1094a 19-21, briefly alludes to a different infinite sequence: "... we do not choose everything for the sake of something else (for at that rate the process would go on to infinity, so that our desire would be empty and vain)." I think this is best read as envisioning an infinite series of means unconnected to any end. In the next section, I discuss somewhat further the argument in which this passage appears.

thinks is worthwhile for its own sake, for that, even if he has no name for what the two jointly constitute, would itself be an end that is not part of the cycle. Perhaps, each of the two is valued both as an end and as a causally necessary (or useful) means, as one may drive, both to reach some destination and also because one enjoys driving. He would still eat if it weren't necessary to work and would still work if it weren't necessary to eat.

Perhaps, this condition in which there are mutually supporting ends which are not viewed as constitutive of any further end can appropriately be described as meeting both conditions, so the agent does have a finite cycle of ends but does not have any final ends. If so, there is still an important point to be made. The kind of life shaped by a finite cycle of ends, even if possible, is not one that can be recommended to anyone whose life cannot *already* be so characterized. No one who does not already guide himself by exactly that cycle of ends will be able to see such a life as desirable.[47] It fails the most basic motivational condition for an account of eudaemonia, that it be such as to engage the beginner as a beginner.[48]

[47] That is, no one who does not already share exactly that cycle of ends will be able to see it as desirable unless he sees that cycle of ends as means to or constitutive of something else that he finds desirable, but, to the extent to which that is true, his ends will differ from those of the agent whose life is shaped entirely by that cycle of ends.

Perhaps, there is a loophole here. Consider a *definitively achievable end*, where that is an end that, upon being achieved, ceases to be an end. My end might be to take a walk. When I have taken the walk, I no longer have that end; it has been definitively achieved. Now, suppose that an agent has a definitively achievable end of coming to have a mutually supporting set of ends, such as eating and working. Then, that agent, whose action is not already shaped entirely by a given cycle of ends, could have a reason for coming to shape his action entirely in terms of some finite cycle of ends. Conceivably, an account of eudaemonia couched in terms of a finite cycle of ends could in this way appear desirable to someone whose actions were not already shaped by that finite cycle. However, it must, in the first place, be judged unlikely that there is anyone with the requisite definitively achievable end, and in the second, the possibility depends upon any such person having a different structure of ends, albeit a structure to be superceded, that does include final ends.

[48] Arguably, it may fail first-personally as well, even for the agent whose cycle of ends it is. If he is sufficiently reflective to imagine possible alternative structures of end-pursuit, he will be able to wonder why he guides his life by just these ends and will not be able to find an answer along the lines that they make up or contribute to a better or more worthwhile life. Should he pose to himself the question why he

4.32 Ultimate Ends

So, it is practically inescapable for beings such as ourselves that we have or at least think in terms of final ends. There are objectives aimed at for their own sake and not for the sake of anything further. Must there also be, for a given agent, an ultimate end, some final end to which all other objectives are related as means?[49] Now, this question might be understood in at least two ways. The question might be whether, as a matter of human psychology, there must be an ultimate end,[50] or it might be whether having or coming to have an ultimate end is normatively necessary.

Aristotle offers an argument that appears to bear on the first question. It occurs in the *Nicomachean Ethics*, I.2:

> If, then, there is some end of the things that we do, which we desire for its own sake (everything else being desired for the sake of this), and if we do not choose everything for the sake of something else (for at that rate the process would go on to infinity, so that our desire would be empty and vain), clearly, this must be the chief good. (1094a 18-22)

On the face of it, the form of this argument is:

(1) If P then Q

guides his action by just this cycle of ends, the most he could say, it appears, is that he just does and sees no reason to change. (See note 42.)

[49] If an agent does have an ultimate end, then that will also be her only final end; she may have other ends, but no others that are final.

[50] Having already addressed this in Chapter One, I shall set aside here the question whether the having of an ultimate end is a matter of human nature in some sense distinct from saying that it is a matter of human psychology.

(2) Not R (because S)

(3) Therefore, Q

That's puzzling unless there is an unstated premise to the effect that R is the only relevant alternative to P. So, filling in, the best reading of the argument appears to be along the following lines:

(1) If there is some end of the things we do, this must be the chief good.

(2) Either there is some end of the things we do or we choose everything for the sake of something else.

(3) We do not choose everything for the sake of something else (for at that rate...)

(4) Therefore, this must be the chief good.[51]

If this reconstruction is correct,[52] then the argument is fallacious. In the sense in which it is plausible that the second premise is true, there being some end of the things we do has to mean that there is some end for each of the things we do – that is, that we do not act without some end or other. That, of course, is consistent with our different actions having different ends in view. But it is not plausible that the first premise is true unless the existence of some end of the things we do means that there is some *single* end of all the things we do. (Otherwise, what does "this" mean in the consequent?) So, if the second premise is true, the antecedent of the first may be false. Since the truth of the

[51] This is based on Ackrill 1980, 25-26. There are further complications in the interpretation of this passage which, not being relevant to my present concerns, I omit.

[52] It is difficult to find a reading that is both illuminating and non-fallacious. In particular, there is a problem understanding the role of the clause denying that we choose everything for the sake of something else. But if that clause is omitted, it is not clear what argument is being offered other than the trivial passage from "if P then Q" to "if P then Q". Additionally, it is plain that Aristotle means to get more than a conditional assertion of the existence of a chief good.

antecedent isn't insured by anything else in the argument, the conclusion that "this must be the chief good" does not follow. This argument, then, does not show that there must, as a matter of human psychology, be an ultimate end.

Nor is it obvious how any other argument could provide support for that conclusion. It might be asserted that if an agent has two ends, then he has a further end constituted by the compound of the two.[53] If that were correct, then, for any agent who has any end at all, there would have to be also an ultimate end (if he has only one end, then the same one). This would make the existence of an ultimate end definitional, and the ultimate end would have no more or different normative force than the ends that constitute it. In terms of a pair of ends, *A* and *B*, when those alone are relevant, we can already say that there is a consideration in favor of any action that advances one without damaging the other, in favor of any action that advances both, against any action that damages both, and against any action that damages one without contributing to the other. But if the presence of some compound end is supposed to follow definitionally from the presence of any other ends, the *A-B* compound does no more than *A* and *B* separately did. For the compound to have any independent normative force, it must be more than just a compound; it must, for example, establish some kind of trade-off or priority relations (at least for some range of cases) that apply when actions in the service of *A* are actions in the disservice of *B*. If the compound has independent normative force, its presence cannot be guaranteed definitionally; if it does not, there is no point in introducing it.

As already mentioned, however, we can understand differently the claim that there

[53] Ackrill suggests, in mitigation of (though not exoneration from) the charge that Aristotle is guilty of a fallacy in the passage quoted above, that Aristotle may have accepted such a premise. (1980, 26)

must be, for each agent, an ultimate end. It might be that one ought to have or come to have some ultimate end. That is a claim any eudaemonist theory is committed to, for eudaemonia is conceived as just such an ultimate end. Accordingly, any eudaemonist is committed to the more modest psychological claim (than that each agent must have an ultimate end) that it is possible for an agent to have or come to have an ultimate end, or at least to approach having one.[54] Second, the eudaemonist, if he does not take the presence of an ultimate end to be guaranteed by the psychology of his addressees, is committed to the normative claim that there are reasons supporting the acquisition of or the approach to having an ultimate end. For the present, I only note that this must be part of the eudaemonist case and suggest that the obvious normative necessity of an ultimate end for a eudaemonist theory may be the reason that some eudaemonists, such as Aristotle, were too quick to conclude that an ultimate end must be present as a matter of psychology.

4.33 Constitutive and External Means

Also important for understanding the eudaemonist classification of ends and means is the distinction between *constitutive means* and what I have called *external means*. The distinction received some discussion in the last chapter and has also been deployed above in the discussion of centralism and reductionism, but it has further-reaching import and will repay more careful consideration. We need to address both the relation of constitutive and external means to one another and the relations in which they may stand to ends – in particular, whether external and constitutive means may themselves be ends.

[54] See note 39.

Intuitively, the distinction is easy to illustrate. I may decide to take exercise for the sake of my health. Here, the exercise is an external means to health. It is something that causally contributes to my health. I may also settle upon tennis as the form of exercise I will take. In some sense, that is taking means to get exercise. Tennis can be compared with other options (volleyball, walking, swimming, thumb-twiddling, etc.) as a better or worse form of exercise. But this is a very different sense of taking means to some objective than is the taking of exercise for the sake of health. The playing of tennis *constitutes* the exercise that I get rather than simply causally contributing to it. Or, here is another example: I may purchase a tie as a means to being well-dressed,[55] but *wearing* a tie is not in the same sense a means to being well-dressed; it is part of what it is to be well-dressed.[56]

On the face of things, we can distinguish the two in terms of whether the relevant objective can be independently specified. When something is an external means to some objective, the objective can, in principle, be fully specified independently of reference to the means.[57] We know what health or being well-dressed is without talking about exercise or the purchasing of ties. A doctor could determine whether I am healthy without knowing what, if any, exercise I engage in. Someone suitably sensitive to the conventions that define being well-dressed could determine whether I am well-dressed without launching any inquiries about where or whether I had purchased a tie. I might be healthy without taking exercise, and a borrowed rather than a purchased tie might

[55] I am speaking of being well-dressed in accordance with certain conventions. That those conventions are not universal does not affect the point.

[56] I owe this example to Roderick Long.

[57] In practice, of course, objectives are rarely, if ever, fully specified.

contribute to my being well-dressed. An external means contributes causally to its objective, and even if it is, in the circumstances, the only way to achieve or promote the objective, it could, if circumstances were different, be replaced by some other means without detriment to the objective.

When, by contrast, something is a constitutive means to some objective, it is not possible, even in principle, to fully specify the objective independently of the means. What the objective is is at least in part constituted or made what it is by the means adopted. The contribution of the means to the end is not, or is not just, causal. If wearing a tie is constitutive of being well-dressed, there is no adequate way of saying what it is to be well-dressed that does not refer to tie-wearing. When a means is constitutive of some objective it stands in some logically or conceptually necessary relation to that objective[58]: Its necessity for the objective is not, as may be the case when an external means is necessary to some objective, a matter of the absence of some other causally effective or useful means to promote or achieve the objective. The constitutive means is necessary for the objective to be what it is.

There are at least two further important points of comparison between external and constitutive means. First, an external means may be sufficient to achieve the relevant objective if enacting or adopting the means is, relative to the situation in which it is adopted, all that is needed to achieve the objective. The parallel, for constitutive means,

[58] The conceptual connection may be very attenuated, as in the relation between playing tennis and taking exercise, since so many different activities, including ones not yet conceived, may be forms of exercise. Nonetheless, it is real; some things, such as taking a nap, cannot count as ways of taking exercise. Also, as exemplified in the case of the various different activities that may constitute taking exercise, the necessity in question may be the necessity for doing *something* that constitutes the objective without, in the absence of further considerations, ruling out the possibility that other things could equally well constitute the objective. The fact that I could swim rather than play tennis as a form of exercise does not imply that playing tennis does not constitute my taking of exercise.

is that some constitutive means may entirely constitute the objective to which it is a means as playing tennis constitutes a way of taking exercise. An external means may also be necessary to its objective if it is not causally possible to achieve the objective without adopting that means. However, an external means need not be either necessary or sufficient for its objective. Taking regular exercise, for example, is not sufficient for health because things can go wrong with health that exercise does not prevent. (Some forms of exercise might even damage a person's health.) Nor is it necessary, for, however unlikely it might be, a person might be in good health without taking any regular exercise. The most that can be said is that regular exercise increases the probability of good health. For a constitutive means, though, matters are different. A constitutive means may not be sufficient for its objective – when it only partially constitutes it, as wearing a tie only partially constitutes being well-dressed – but it is always necessary: It does not *merely* increase the probability of its objective. Of course, wearing a tie *does* increase the probability of my being well-dressed, but it increases it from zero – the probability that I will be well-dressed without a tie – to something greater than zero, not from a lesser to a greater positive value.[59]

This is connected to a second important point, that external means admit of trade-offs or substitution[60] in a way that constitutive means do not. If my objective is to begin to invest a certain sum every month, then, in order to have that sum available each month, I may either reduce other expenditures or attempt to increase my income. Either of these

[59] See the qualification in note 58. What is necessary may be that some constitutive means or other be adopted.

[60] There is not a sharp distinction between substitution and trade-off: Substitution is just the limiting case of trade-off.

courses of action is an external means to having the sum available for investment. So far as that is the only relevant objective, one may be substituted for the other or they may be combined in various ways. I can compensate for not having increased my income sufficiently by reducing expenditures elsewhere. For constitutive means, this is not the case. If, to use an earlier example, wearing a tie and wearing a belt are both constitutive of being well-dressed, then I cannot make up for not wearing a belt by wearing a nicer tie (substitution fails), nor can I compensate for wearing an ugly tie by wearing a nicer belt (trade-offs are not possible).[61, 62]

Now that we have a better sense of the contours of the distinction between external and constitutive means, we can begin to address the relation that either may have to ends. It is obvious that neither can be a final end; being any kind of means precludes that, for a final end is one that is not sought, pursued or performed for the sake of something else.

It is almost as obvious that either can be an end. As an example of an external means that is also an end, I may mention again the case of driving to reach a destination, both in order to reach the destination and because one enjoys driving for its own sake. Driving is an external means to reaching the destination; we can certainly understand

[61] This does not imply that no comparisons are possible. Of two beltless persons, one might be better dressed because he is wearing a nicer tie, though it would be improper to say of either, without qualification, that he is well-dressed. Or, though again neither could be said without qualification to be well-dressed, one who wears no belt and a nice tie might be better dressed than someone who wears a nice belt and no tie, because a tie is a more prominent constituent of attire.

[62] There is additional complexity which, though it does not require any alteration of the analytical points already made, is worth mentioning. Just as external means can be arranged in series or chains, with a given means becoming in its turn an objective to which further means are sought, so constitutive means can be nested within one another. Playing tennis may be constitutive of the exercise I take, and lobbing the ball across the net constitutive of the tennis I play. Moreover, constitutive means may serve an external means as their objective, as in the selection of tennis as the form that my exercise, itself an external means adopted for the sake of health, takes, and constitutive means may also serve as objectives to which further external means are anchored, as when I reserve a court to play tennis.

what it is to be at the destination without knowing or referring to the way in which one arrived there, and there may be alternative ways to reach the destination, such as taking a bus. Still, driving may not be selected over its alternatives solely because it is more efficient or faster or the like (even if it is), but because it is enjoyed for its own sake.

Constitutive means can also themselves be ends. Playing tennis may constitute the exercise I take, but may also be or come to be found worthwhile for its own sake. Though it may be that I would not have taken up tennis except as a form of exercise or determined upon exercise except insofar as it was expected to contribute to my health, it may be that in the playing of tennis, I come to enjoy the game itself in addition to caring about the health-related benefits. If, for example, some study were to show that tennis, contrary to prior opinion, had no significant positive impact upon health, I would not then give it up. For, though health benefits were the initial reason for taking up the game, they are not the only reason for continuing to play.

So, either external or constitutive means may be ends. As means, they may be related to objectives that are themselves ends or to objectives that are not ends. Plainly, there is nothing about an external means, simply insofar as it is an external means, that requires that it also be an end, whether or not its objective is an end.[63] There are more interesting questions, however, connected with whether constitutive means may or must be ends. It is also fairly obvious that a constitutive means may be an end when the objective of which it is constitutive is itself only an external means (and not also an end) to something further. I may take a job solely because of the way the particular job satisfies various parametric conditions such as salary. Had some other position offered a

[63] An external means *may* be an end, though, even if its objective is not an end.

better salary, while not being worse in terms of other conditions (such as proximity to my residence), I would have taken it instead. Taking the job is an external means to receiving a certain salary. Once I have taken the position, though, I may find that something constitutive of the job I am expected to perform is itself something I find worthwhile for its own sake.[64]

But suppose that the objective of which the means is constitutive is itself an end. Must the constitutive means *then* also be an end? (Call this the Constitutive End Thesis.) I think the Constitutive End Thesis is initially plausible, but I do not think it is entirely obvious. It is uncontroversial and obvious that such a constitutive means *may* be an end. I shall quickly, and without discussion, give two examples to make that point. At greater length, I shall examine three possible counter-examples, and consider whether they are genuine.

It is not difficult to present examples in which a constitutive means to something aimed at or performed for its own sake *does* function as an end also. There is, say, playing tennis well, enjoyed for its own sake, and there is also, constitutive of it, the gracefully executed return. Or there is a friendship, sustained for its own sake, and there are also, constitutive of it, various shared activities, such as conversation.

It is not so easy to find plausible examples in which a constitutive means to something aimed at or performed as an end is not itself an end. But consider again the case of a genuine and close friendship, sustained for its own sake. Part of what is essential to such friendship is that each friend care about the other's well-being for its

[64] I might even have foreseen that this would be the case, so long as having foreseen it did not play a role in the decision. Also fairly plainly, when the objective is not an end, it is possible for the constitutive means not to be an end either. Whatever job I take, there will be something or other constitutive of what I am expected to perform, and I may not find that worthwhile for its own sake.

own sake. What happens when one friend learns that disaster or serious harm has befallen the other? No doubt, she will do what she can to help. But also, she will feel sorrow or grief over what has happened. This sorrow is not just some external means or accidental accompaniment of her concern for her friend. It is essential to and constitutive of the concern for a friend's welfare that is part of what it is to be a genuine friend. If she did not feel the sorrow, she would not be a genuine friend. Surely, though, it might be argued, that does not mean that she values the sorrow or grief for its own sake.

Even on its own terms, I do not think this is entirely clear. To say that something is an end is neither equivalent to nor does it imply that it has a certain affective quality.[65] And the grieving friend may well say that she of course does not *enjoy* sorrow or grieving, but that it is not something she would do without, even if she could. But though I think this may have considerable merit, I will not pursue this line of response. Instead, I will point to a distinction that applies even if the case cannot be made that sorrow over harm to one's friend is an end.

The key distinction is between what is constitutive of an end and the special case in which something is a constitutive *means* to an end. Whenever there is an end, or more generally any objective, there is some state of affairs (which is not normally fully specified) that is conceived to be possible or at least possible to approach. Of course, there must be *something* that constitutes this state of affairs. But constitutive means to an objective need only partially constitute it. There may be other features of the envisioned state of affairs that also partially constitute it without being constitutive means.

[65] Certain affective qualities may, however, make it extremely unlikely for something to be adopted or pursued as an end.

This distinction provides the needed tool for considering sorrow as constitutive of concern for a friend's well-being. The first point to notice is that it is misleading to speak, as I just did and as in the initial description of the case, of the friend's sorrow as constitutive of her concern for another's well-being.[66] The right way to describe the case is that being such as to feel sorrow or grief at harms to the well-being of a friend – that is, having a certain dispositional state of character – is constitutive of being concerned with the friend's well-being. But this, being such as to be grieved at serious harm to a friend's well-being, is not a constitutive *means* to having the concern. One does not take it as one's objective to come to have that dispositional feature in order to care for a friend's well-being. It is rather that, *in* caring for or coming to care for the well-being of another for its own sake, one is such or comes to be such that one would feel sorrow at harm to the other.

Consider a different and more difficult case. Suppose that one considers one's job worthwhile for its own sake. One is engaged in work that one considers important and valuable. But constitutive of performance of the job is dealing, sometimes, with a corporate bureaucracy. That is indeed a means – one must deal with the bureaucracy in order to perform the job – and it is indeed constitutive of the job one has – the job would not be the same if dealing with the bureaucracy were not required. But it seems intelligible that dealing with the bureaucracy is not among one's ends; it is, from the standpoint of the job-holder, an unfortunate concomitant.

Since dealing with the bureaucracy is undeniably both constitutive of and a means

[66] It is doubtful that it is intelligible at all, and if it is intelligible, it sounds nasty. (If my sorrow is constitutive of my aiming at your well-being, then, in aiming at your well-being, I aim at my own sorrow. But since my sorrow is at harms to your well-being, then, in aiming at your well-being, I aim for you to be harmed.)

to performing the job, there are only two possible tacks for reconciling the case with the Constitutive End Thesis. It might be that dealing with the bureaucracy really is an end or else that performing the job is not.

The more plausible is that performing the job is not an end. It might be said that there is some aspect or component of doing the job, such as creative work involved, that is found worthwhile for its own sake and that, though initially one might have been inclined to say that it was the performance of the job itself that was valued as an end, on further reflection, so describing the case is a misreading of the relevant motivations. Instead, though both the creative work and dealing with the bureaucracy are constitutive of the job, dealing with the bureaucracy is not constitutive of the creative work, and it is only the creative work that is performed for its own sake.

This seems possible, but it is not enough to make the case for compatibility with the Constitutive End Thesis unless we can rule out the alternative hypothesis that it really is the job that is taken to be an end, but that its being an end is not a simple function of and does not imply that its constituents, considered apart from their place in the end they constitute, are ends. The status of the job performance as an end is, for the agent, a property of its constituents standing in certain relations to one another. I do not see how this can be ruled out. It is, for example, arguably a familiar feature of works of art that they have or are taken to have value as wholes (often referred to as a matter of their organic unity) that is not reducible to (say) the value of the separate brush-strokes in a painting or the notes in a musical composition. It is how the parts are put together that is at least partially responsible for the value of the whole, not the value of the parts considered separately. This is, perhaps, only an analogy, but I do not see how to argue that ends and the value of works of art must be disanalogous in this respect.

The other attempt at reconciliation with the Constitutive End Thesis involves the claim that performance of the job was an end and that dealing with the bureaucracy, as a constitutive of job performance, is also an end. This seems much less plausible. For, to treat something that is a constitutive means as an end also is to imply that one would, for at least some relevant range of counterfactual cases, still perform it if it were no longer necessary to or constitutive of some other end that one had. But surely it is possible that one would have not even the slightest disposition to deal with the corporate bureaucracy if it were not part of the job. And if that is so, then dealing with the bureaucracy is not an end, and so, the analysis does not effect a satisfactory reconciliation of the case with the Constitutive End Thesis.[67]

Consider a third case.[68] Suppose that one's end is to play the violin. Constitutive of that and means to it are both using one arm to hold the violin against one's chin and making certain movements with a bow with the other hand. But neither of these alone are ends. The violin-player would not recognize anything worthwhile, certainly not anything worthwhile for its own sake, in holding the violin against his chin without playing it or in making various stroking motions with the bow if no violin were present to stroke.

Perhaps, this case can be reconciled with the Constitutive End Thesis, but it looks unlikely to the point of desperation. The same kinds of options as were available in the case of dealing with a corporate bureaucracy are available here. Since stroking (with the

[67] As applied to the present case, the Constitutive End Thesis implies that the conditional, "if job-performance is an end, then so is dealing with the bureaucracy," always holds. To deny that requires that there be some possible case in which job-performance is an end, but dealing with the bureaucracy is not. The considerations of the last two paragraphs are, in different ways, meant to suggest that this is a real possibility.

[68] This was suggested by Fred Miller.

bow) and holding (the violin) are both constitutive of and means to playing the violin, a friend of the Constitutive End Thesis would have to maintain either that playing the violin is not an end or that stroking and holding are ends.

The problem with the former is that there does not seem to be anything into which playing the violin can be decomposed (as job-performance could be decomposed into creative work and dealing with the corporate bureaucracy), so that some part or aspect of the violin-playing can be regarded as the end, while the stroking and holding can be viewed as constitutive means to some other part or aspect of violin-playing which is not an end. The problem with the latter is that there appears to be no relevant range of counterfactual cases such that one would still hold the violin or still stroke with the bow if they were not constitutive of violin-playing.[69] There may be some way to avoid these conclusions and thus maintain the Constitutive End Thesis, perhaps by way of some account of how acts can properly be individuated, but unless that is further spelled out and defended, I think the case of holding and stroking as constitutive means to violin-playing has to be accepted as a genuine counter-example to the Constitutive End Thesis.

With cases of this sort in mind, it seems to me that it is at most *barely possible* that the Constitutive End Thesis is true. But so long as we have only such examples and analyses upon which to base a judgment, I do not see that we can say more for it. The general form of any proposed counter-example would be that there is some plausible case in which a means constitutive of an end is not itself an end, and the general form of any

[69] One might do one or the other by itself in the course of practice – say, to accustom oneself to the way the violin feels cradled in one's arm or to the range of motions required to draw the bow across the strings – but that does not seem to be a relevant case in which one would adopt the means apart from their contribution to the end of violin-playing; instead, these would be activities adopted only for their expected contribution (as external means) to violin-playing.

alternative analysis offered in reply would be that there is some *more* plausible reading of the case such that it is not a genuine counter-example.[70] There seems to be no reason to think this will always be true, and for some cases, such as violin-playing, it appears to be false, so it will be hard to construct any general argument that the comparative plausibility judgments invoked will always favor the thesis. If we work only with alleged counter-examples and their respective analyses, we will have to proceed case by case, and the thesis will be more or less credible depending on the outcome of the particular analyses. For present purposes, I think that we cannot confidently affirm the Constitutive End Thesis; only the more limited claim, that constitutive means to an end *may* themselves be ends, is warranted.

However, this is not quite the end of the matter. There is also to be considered the special case of constitutive means to an end in which the end of which the means are constitutive is a final end. Here, I think we can argue for what might be called the Restricted Constitutive End Thesis, that a constitutive means to a final end must itself be an end. For what made it possible to maintain that a constitutive means to an end might not itself be an end was that if the end could be altered so as to remain the same but for that constitutive means, there might be reasons for altering it – that is, for replacing it with the altered end – in terms of other ends (such as avoiding distasteful activities like dealing with a corporate bureaucracy). But if the end of which the means is constitutive is final, then it is sought, aimed at or performed for its own sake and not (at all) for the sake of anything else, so there are no further ends which bear on the acceptability of

[70] Counter-examples might fail to be genuine along several dimensions – the alleged constitutive means to an end might not be constitutive of an *end*, might not be constitutive, might not be a means or might, despite appearances, be an end.

pursuing or aiming at it. One could not have reason to alter a final end so as to omit or replace something constitutive of it unless it were no longer final.

4.4 The Virtues: Their Place within Eudaemonism

For eudaemonism, the moral virtues or excellences[71] are pivotal both to understanding what it is to live well and to actually living well. Traits of character such as honesty, loyalty, fairness, compassion, generosity, conscientiousness, tolerance, kindness, courage and, most generally and importantly, practical wisdom[72] are invoked as essential to living well[73]:

> For no one would maintain that he is happy who has not in him a particle of courage or temperance or justice or practical wisdom, who is afraid of every insect which flutters past him, and will commit any crime, however

[71] There are also, according to Aristotle, intellectual virtues or excellences. In outline, the distinction works like this. The human soul – roughly, the functional organization of a human being – is divided into several parts. There are both rational and non-rational parts. Among the non-rational parts are the desiring and appetitive functions. Though they are non-rational (perhaps *inarticulate* captures much of the sense, here), they respond to reason and have "a tendency to obey [it] as one does one's father." (*NE* 1103a 2-4). The moral virtues or excellences of character are properties of this non-rational part when it is disposed to behave and respond rightly to the situations with which a person is confronted. By contrast, the intellectual virtues are excellences of the rational parts of the soul, but apart from *phronesis* or practical wisdom, which is classed with the intellectual virtues, I shall not be concerned with them.

[72] Practical wisdom differs from other virtues in that the other virtues involve appropriate action and responsiveness to particular situation-types – honesty being correlated with communicative situations, courage with dangerous situations, and so on. But practical wisdom is, so to speak, a master-virtue, having to do with when honesty or courage or something else is called for and also qualifying the claims of each in light of whatever other virtues or other considerations may be relevant. The practically wise person is, quite generally, responsive to whatever is relevant to right action in a particular situation and prepared to act accordingly. It is only through the inclusion of practical wisdom in the catalogue of the virtues that it is plausible that the person who has all the virtues will always act and respond properly.

[73] The list of virtues of course differs somewhat in different thinkers and as conceived in different, socially embedded, traditions. For the present, I am concerned with issues that abstract from these differences – with the way in which the virtues fit into eudaemonism, not with the threat of relativism that the differences might be thought to present. Put differently, I am concerned here with the concept of virtue rather than with different conceptions of virtue. (On the concept-conception distinction, see Rawls 1971, 5 and Dworkin 1978, 134-136.)

> great, in order to gratify his lust for meat or drink, who will sacrifice his dearest friend for the sake of half a farthing, and is as feeble and false in mind as a child or a madman. These propositions are almost universally acknowledged as soon as they are uttered ... (*Politics* 1323a 26-35)

The first, and in one sense the easiest, issue to address is the question as to the sense in which the virtues are essential to living well. The virtues are, at least partially, constitutive of the ultimate end, eudaemonia, and as such are ends themselves and must be practiced for their own sakes, not just for the sake of something further to which they contribute. With the possible exception of the Epicureans,[74] this is the uniform position of the ancient eudaemonists. At one extreme, the Stoics held that possession of the virtues was both necessary and sufficient for eudaemonia, that they entirely constituted it. But more moderate positions, such as Aristotle's, while denying that the virtues alone were always sufficient to live well, also held that they were necessary and practiced for their own sakes. For "good action itself is its end", (*NE* 1140b 6) and

> The agent also must be in a certain condition when he [performs acts that are in accordance with the excellences]; in the first place he must have knowledge, secondly he must choose the acts and choose them for their own sakes, and thirdly his action must proceed from a firm and unchangeable character. (*NE* 1105a 28-1105b 1)

That is not only the traditional answer but also, I think, the most defensible in its

[74] See note 2.

own right.[75] And though this goes some way toward saying what the place of the virtues within eudaemonism is, to leave matters at this would omit a great deal that is important both for understanding the virtues and for understanding what eudaemonia is taken to be. For the present, I shall highlight one further feature of the virtues as they figure within eudaemonism, their pedagogical role.[76]

Whatever a virtue is, it is, as indicated in the quote above, a stable trait of character. A person is not, for example, honest just because she tells the truth on some occasion, not even if the occasion is one upon which it is for some reason tempting not to be truthful. (It may, of course, be good evidence for her honesty.) To say that she is honest is to say at least that her character is such that she could be expected to be truthful in some class of cases in which the situation calls for it.[77] A virtue involves acting in a certain way, being intelligently responsive to the situation that calls for that kind of action, and being motivated and feeling appropriately.[78] As stable traits of character, the virtues, taken separately, can be recognized in others (and sometimes ourselves), and are admired and praised in those who possess them. Taken together, the practice of the virtues (not necessarily limited to the list above) comprises the ways in which a morally good person acts, responds and is motivated in the issues and situations with which she is concerned.

[75] It is not my purpose, however, to defend it here, just to present it as part of the eudaemonist position.

[76] There is, of course, much more to be said about the virtues, some of which is addressed in Chapter Five.

[77] The situation may not call for it. She is not dishonest, for example, to abstain from giving a full medical report in response to a casual question about her health.

[78] This presupposes that our emotional responses are, at least in part, cognitive and therefore educable. For discussion, see Solomon 1976 and Nussbaum 1994 and 2001.

That a virtue is recognizable presupposes at least some body of discourse or linguistic practice which picks out, on the one hand, some range or class of cases, and on the other, a certain kind of response thought appropriate to cases within that class. This recognizability of a virtue allows it to play a kind of pedagogical role within a eudaemonist theory.

There are several related points here. First, as discussed above, eudaemonia is introduced as a thin concept, a concept of something that will satisfactorily answer to concerns about what it is to live well. To be embodied in practice, it stands in need of specification; we need to say what it is concretely to live well. The virtues provide a beginning in that direction. In terms of recognized virtues, it can be said that *this* – being honest in communicative situations, courageous in the face of danger, generous when the wants or needs of others can be met at modest cost[79] and so on – is what is involved in or required by living well.

Second and also important, the specification does not have to be limited to some verbal formulation. For the particular virtues, there are models or exemplars who already embody the character traits in question. The beginner who, for the first time explicitly and with practical intent, is approaching the question of how to live well, can be pointed to those who exemplify a virtue, to those who are, for example, honest or courageous or generous. The availability of such exemplars has the two further functions of providing proof by example that the recommended traits of character are possible, that they can be acquired and embodied in one's life, and of providing occasion for assessment as to

[79] If and when meeting needs is required by justice is, of course, a different matter, not within the province of generosity.

whether and how the traits really do fit into and are required by a desirable life. Theoretical arguments about the content and desirability of the virtues can only go so far; practical demonstration can go much further.[80]

Third, though a virtue is not just a habit, it is still true that "moral excellence comes about as a result of habit.... we become just by doing just acts, temperate by doing temperate acts, brave by doing brave acts" (*NE* 1103a 16, 34-1103b 1). This provides at least the beginnings of the developmental account that the eudaemonist needs.[81] In recognizing that virtue is formed through habit, steps that can be taken by the beginner towards becoming virtuous and thus towards living well can be identified. As she learns, her responses will no doubt be refined and become more sensitive to the important features of the situations in which the practice of a virtue is called for, but refinement presupposes something to be refined and, to be of use for the developmental account, something that is immediately, without further preparation, accessible to the beginner.

4.5 Summary

For eudaemonism, the central moral conception is living well or having a good life. This is easily misunderstood in at least three ways: by identifying eudaemonia with happiness, construed as a subjective or purely experiential state, by taking 'living well' to mean 'having a life that serves one's interests,' and by trying to reduce all other moral or normative conceptions to their role as contributions to eudaemonia, conceived as an end

[80] One advantage, on the score of realism of assessments, is that, when a virtue is actually embodied in persons' characters, it is subject to all the shocks, surprises and unforeseen consequences that the world can throw at it, but not to those that occur only in thought-experiments.

[81] I do not, of course, mean that habituation is all that is needed.

capable of being understood independently.

When these misunderstandings are avoided, we are in a position to see that eudaemonia is introduced as a thin term, a place-holder for something that will satisfactorily answer to concerns about what it is to live well, that something that will answer to those concerns amounts to an inclusive ultimate end for action, and that it stands in need of further specification. Further specifying what eudaemonia is involves at least three things. First, there are formal constraints upon the notion that must be satisfied by anything that might count as eudaemonia. Second, there are developmental and motivational constraints, for coming to have eudaemonia as one's ultimate end (or to approach doing so) must, if it is to answer to the concerns that motivate the search for an acceptable conception of eudaemonia, be something that the searcher can see as answering to those concerns. Third, since eudaemonist theories characteristically account for their prescriptions in terms of ends recommended and what is thought to contribute to those ends, a framework for thinking about the relations and inter-relations of means and ends, and about the different types of means and ends there can be, is needed in order to avoid over-simplifying and therefore misrepresenting the kinds of considerations that can properly figure in deliberation.

In terms of this kind of (still very abstract) account of what eudaemonism is and involves, and especially by relying on the discussion of the relations of ends and means, we can see what kind of place the virtues have within the eudaemonist framework. They are, first, constitutive means to eudaemonia, practiced both for their own sake and because they are necessary to eudaemonia, and second, they figure in the pedagogical and developmental story upon which eudaemonist theories must rely.

CHAPTER FIVE: REASONING ABOUT ENDS

5.0 Introduction

Can we bring instrumental reasoning to bear on the selection of ends, and especially upon the selection of final or ultimate ends? Widely shared assumptions suggest not. Remember that, in the terminology introduced in the last chapter, an *end* is an objective sought or aimed for its own sake, that a *final end* is one that is not sought or aimed at for the sake of any further objective, and that ultimate ends are a subset of final ends. Those facts, in combination with the uncontroversial premise that instrumental reasoning consists of adapting means to ends, make it natural to infer that instrumental reasoning can have nothing to say about final ends. However far instrumental reasoning can reach, it will have to proceed in terms of some further end. No end or objective to which some tract of instrumental reasoning leads can be final, for anything to which it leads will be adopted or selected for its conduciveness to the further end. Accordingly, if there is any bearing of practical reason upon final or ultimate ends, it must be along some non-instrumental route.

Matters look more desperate yet – so far as the rationality of final ends goes – for those who also endorse *instrumentalism*, which can be defined in terms of two theses:

(1) Rational agents have reason to adopt means to their ends. This is the instrumental function of practical reason.

(2) There is no other function of practical reason than regimenting means in terms of ends. In particular, there is no non-instrumental function of reason by which ends are determined or picked out.

Instrumentalists can be expected to accept the argument that instrumental reason cannot bear upon ends, but will deny that practical reason affords us any non-instrumental routes to the selection or identification of final ends. What cannot be supplied by instrumental reason cannot be supplied by reason at all, but must have some non-rational source.[1]

Despite the naturalness of the inferences, I think them mistaken. Instrumental reasoning can bear upon ends, including final and ultimate ends. Reasoning about final ends is open even to the theorist who confines his account of practical reason within the strictures of instrumentalism. Drawing together strands of argument from earlier chapters, what I shall try to do is, first, show that this is possible and second, that it is plausible that the resulting structure of ends has intriguing parallels to what is

[1] To say that the selection of an end has some non-rational source is not, of course, to say that its selection or pursuit is *irrational*, for that would require that there be some way in which practical reason does bear upon ends, by ruling out some and presumably not others as irrational.

recommended in eudaemonist theories. We shall find that instrumental reasoning, proceeding from a normal set of motivations, can lead to an over-arching ultimate end, including within itself other ends, sought or pursued for their own sakes, and that among these are practices expressive of enduring traits of character.

However successfully that project may be carried out, there remain loose ends. In the remainder of the chapter, I shall try to briefly address some of them, with the aim, not of resolving them, but of marking out areas and directions for further exploration.

5.1 Instrumental Reasoning About Final Ends

If instrumental reasoning can bear upon final ends, there must be some defect in the argument against such bearing sketched above. In fact, there are at least two. The first is a confusion of form with content. The second comes from overlooking a particular class of ends.

Suppose that some means, M, is adopted because it promotes an end, $E_{1.}$ Schematically, something like that will be true whenever instrumental reasoning supports the adoption of some means. However, it tells you nothing about the content of the means, M. It may be that M, which genuinely does promote E_1, consists in the pursuit of E_2 for its own sake and not for the sake of anything else, that is, as a final end. That will be possible in principle when the pursuit of E_2 as a final end is sufficiently conducive to E_1.[2] The end, E_1, provides a reason for the adoption of the means, but that does not imply that once the adoption has occurred, the means itself will consist, wholly or partly, of

[2] I mean that the pursuit of E_2 as a final end must be conducive enough to E_1 that its adoption as a means to E_1 is rational. That would generally be untrue if there were markedly superior ways of achieving or promoting E_1.

something done for the sake of E_1. The possibility that it will not is overlooked due to the tacit assumption that the form of an instrumental argument, that the means is adopted for the sake of an end, dictates the content of the means adopted.

A second reason that the possibility of instrumental reasoning to a final end is overlooked is through failure to consider a class of ends from which the reasoning might proceed. Some ends organize action on an ongoing basis – health or wealth, for example. Others, however, may be definitively achieved and, when achieved, no longer function as ends. If the end is to take a walk, then, once the walk is taken, one no longer has that end. Overlooking definitively achievable ends may contribute to failure to recognize the possibility of instrumental reasoning to a final end. Simply put, adopting some final end, E_2, may achieve what is aimed at in some definitively achievable end, E_1. If so, E_2 will remain to direct later action, while E_1, in the service of which it was selected, vanishes from the agent's body of ends. E_2 will not, on an ongoing basis, be directed to the service of E_1 because E_1, having been definitively achieved, is no longer an end.

In principle, then, instrumental reasoning can bear upon final ends and thereby at least potentially upon ultimate ends. So far, however, that is only in principle. The fact that we can conceive of instrumental reasoning leading to the selection of a final end does not tell us that there are any interesting cases in which it does. For that, we need to go beyond any purely formal approach.

5.11 Schmidtz's Maieutic Objectives

David Schmidtz has noted the same two points, that there is a "distinction between *pursuing* a final end (which by definition we do for its own sake) and *choosing* a final end (which we might do for various reasons),"[3] and that there is a possibility of "eliminating [an] earlier goal as an item to pursue [by achieving it]."[4]

These points are exploited to call attention to the existence of *maieutic objectives*,[5] which are "achieved through a process of coming to have other [objectives]."[6] Among the plausible examples he offers are the goals of settling upon a career, of selecting a spouse, and of finding something to live for. One seeks to settle upon a career only until one has done so, to select a spouse only until a suitable and willing spouse has been selected, to find something to live for only until something has been found. In each case, the point is that the initial goal, the maieutic objective, no longer structures or guides action but is replaced by something appropriate to what has been settled upon. One seeks to do well in the chosen career, to live happily with one's spouse, to promote the cause or causes one has selected.

Schmidtz incorporates these themes into a series of models for the structure of a person's ends and concludes that it is possible, in principle, for there to be no "loose ends," ends which are simply given, but which are not in any way the product of rational deliberation. Though ends which are simply given are necessary to get the deliberative process underway, it need not *remain* the case that any ends are simply given: every end

[3] Schmidtz 1995, 61. In Schmidtz's usage, a "final end" is just what I call an "end," but the point remains that pursuing something for its own sake is compatible with its pursuit having been selected for some other purpose.

[4] Schmidtz, 1995, 64.

[5] Schmidtz refers to "maieutic ends" rather than "maieutic objectives." This is because, generally, he uses "end" to refer to what I term "objectives." There is no implication in his usage than an end is sought or pursued for its own sake, and thus none that a maieutic end (his usage) must be pursued for its own sake. I have adjusted his terminology to match mine by speaking consistently of "maieutic objectives," which may or may not be aimed at for their own sakes, rather than of "maieutic ends."

[6] Schmidtz 1995, 61.

can be the object of a choice which is both rational *and* instrumental.[7]

So far, this seems right. I have no quarrels with Schmidtz over either the existence of maieutic objectives or the use to which he puts them. What I am concerned to do is not to show that he is wrong, though there are, no doubt, matters in the neighborhood on which we would differ, but to go beyond him in certain respects. What he seeks to show is that there are maieutic objectives that may lead to the instrumental rationalization of a system of ends. For a particular person, the needed maieutic objectives may be present or they may not. Schmidtz believes that the maieutic objective of finding something to live for will often be present in contemporary circumstances, but I think he would agree that it may not be, and, when it is not, no further instrumental argument (on that subject) can be addressed to the person who sees no need to find something to live for.

If, for the moment, we take "finding something to live for" to stand in for coming to be motivated to some degree by moral ideals,[8] shaping one's actions on the basis of some conception of what is right or good or admirable, as distinct from what is effective or efficient with respect to some given set of ends, then what I wish to show is that moral motivation, or the kind of reasoning that can lead to moral motivation is not that contingent. A very general problem, faced by almost all, generates the need to think

[7] Schmidtz 1995, 69-79.

[8] Finding something to live for is 'standing in' for coming to be morally motivated because the kind of moral motivation I have in mind is not readily expressed, without further elaboration, in Schmidtz's terminology. Since he refers to all objectives (my terminology) as 'ends', and uses 'final ends' to denote those objectives sought for their own sake, it is more awkwàrd to express the ideas of (a) objectives sought both for their own sake and for the sake of something else (though he acknowledges their existence – "[a]n end could be final ... and at the same time could be instrumental, pursued as a means to some further end" [Schmidtz 1995, 66]), (b) objectives sought for their own sake and not for the sake of anything else ('final ends', in my terminology), and (c) an objective sought for its own sake, not for the sake of anything else, and to which all other objectives within the system of ends and means which shapes a person's actions bear the relation of being means (an 'ultimate end', in my terminology).

about what to live for, what one's life will be about. I shall try to approach it from two directions, first, by saying something about what it is for there to be something one's life is about, and second, by exhibiting the problem, and the reasoning from the problem, that leads to there being something one's life is about in that sense.

5.2 An Ultimate End: The Shape of a Life

An ultimate end, in the system of a person's ends, is the final end to which all other objectives are means, whether external or constitutive, and whether themselves sought or pursued for their own sakes or not

Consider an *ideally structured ultimate end*. An ideally structured ultimate end would establish trade-off or priority relations among all of the more particular ends that constitute or contribute to it for any situation the agent might face. In its terms, it would be possible to provide an answer as to what to do, what is most important, what is most worth seeking, having or risking and the like, when the decision must be made under some degree of ignorance. In all the situations of a life, guidance could be found in an ideally structured ultimate end.

Even this is not sufficient for such an ultimate end. Providing some decision principle or other is not enough. If all that is wanted is a decision principle to arbitrate between possible conflicts or tensions between ends, that can easily be provided. For example, we could assign importance on the basis of an alphabetical ordering. An ideally structured ultimate end would have at least two further features. It would be embodied in the agent's motivations so that she would actually decide, and view it as reasonable to decide, in terms of the rankings that it generated. Second, it would be reflectively stable

– that is, it would not be subject to being (reasonably) altered or revised in the light of further experience or reflection.

Clearly, if one had an ideally structured ultimate end, that would be enough to say that one had an ultimate end. Equally clearly, no one has an ideally structured ultimate end, if only because the world can surprise us in ways for which we are unprepared by any prior thought or experience.[9] We may be called upon to choose between options that we never thought of as being in conflict, never ranked or prioritized with respect to one another. An ideally structured ultimate end sets, so to speak, the Platonic ideal for an ultimate end compared to which all actual ends fall short.

But that we cannot have an ultimate end in that sense does not imply that we cannot have ultimate ends at all. It implies *either* that we cannot, and therefore do not, have any ultimate ends *or* that something that falls short of being an ideally structured ultimate end may still count as an ultimate end.

I argued earlier[10] that one does not have an ultimate end who merely has two or more separate ends. To make sense of saying a person has an ultimate end, it has to *do* something, has to make a difference to what the person decides or would decide. Thus, the behavior of a person with an ultimate end cannot be explained equally well in terms of the ends that are said to constitute her ultimate end, operating separately. Specifically, I argued that the ultimate end would have to establish some trade-off or priority relations among the separate ends (I shall abbreviate by calling these 'priority relations').

However, having established some priority relations among separate ends is still

[9] The reasons discussed in Chapter Two for denying that we can have a complete preference ordering are also pertinent.

[10] In Chapter Four, § 4.32.

not sufficient for a person to have an ultimate end. For it may be that priority relations have only been established among various subsets of her total system of ends, that there are, so to speak, 'islands' of coherence and prioritized relations among certain of her ends that amount to final ends to which all of *their* constituent or contributing ends are means, but that there is no over-arching end with respect to which all other ends are means. There may, for example, be priority relations established among her ends, *A*, *B* and *C*, and also among her ends, *D*, *E* and *F*. Those priority relations may be sufficient for the existence of a final end, *G*, unifying *A*, *B* and *C*, and for the existence of a different final end, *H*, unifying *D*, *E* and *F*, but so long as she has two or more final ends (*G* and *H*), she can have no ultimate end. In other words, the question whether a person has an ultimate end may re-emerge with respect to ends which themselves include others,[11] and again, there is no point in talking about an ultimate end unless it makes a difference, unless what the person would do, given the ultimate end, differs from what she would do, given the more particular ends alone (though the particular ends we are here considering themselves embody priority relations among other ends that may, partially or entirely, constitute them).

The priority relations among the person's ends must go beyond the establishment of local islands of coherence among her ends. She must also recognize the relevance of considerations of ordering and harmonization among her existing ends and be open to the possibility that the set of ends will stand in need of modification, revision or alterations in the relative importance of its members. The modification or revision called for may

[11] Though the locution of 'one end including others' would most naturally be understood to refer to a constitutive relation between the included and the including ends, I use it here to cover both ways, the external and the constitutive, that one end may be a means to another.

include taking steps to acquire an end that one does not already have, if this promises to better integrate her other ends or taking steps to eliminate some end if its pursuit interferes too greatly with others. It may be better to call this being on the way to having an ultimate end rather than having one, and, when I have occasion to refer to this fact, I shall speak of the *developmental process* involved,[12] but in one important respect, it doesn't much matter. Whether conceived as having an ultimate end or as a process the ideal terminus of which[13] is having an ultimate end, it establishes a dimension along which improvements in the entire system of one's ends can be assessed, and thus does something that none of the particular ends by themselves, nor all of them together, considered only as a collection of ends, could do. It makes a difference to what she has and recognizes having reason to do. I shall abbreviate by calling action in accordance with reasons of this kind, which derive from having the ultimate end or undergoing the developmental process, *action according to the end* or *in terms of the end.*[14]

Is anything more required? At least this much, I think. We also need to insist on more modest analogues of the motivational and reflective stability requirements

[12] Cf. the discussions in Chapter Four, §§ 4.2 and 4.4, and also, in the same chapter, note 39.

[13] That ideal terminus, of course, may never actually be reached.

[14] Should we *always* speak of the developmental process rather than of the ultimate end, on the ground that nothing short of the unattainable ideally structured ultimate end is beyond conceivable improvement, nor, therefore, beyond alteration? I do not see that that is required, for it is relevant *how* the end is supposed to change. To use a standard example, your end may be to have an entertaining evening, and varying possibilities for entertainment may present themselves – a concert or a play, for example. Even after the concert has been selected, new information (the play got a five-star rating from your favorite reviewer) may lead you to alter your plans, but that does not necessarily mean that your end has changed. What has changed is the way in which, rather than whether, you plan to be entertained. Going to a concert or to a play are competing specifications of the end of having an entertaining evening. It would be neither appropriate nor necessary to speak of a change of end unless something else, such as preparing for tomorrow's meeting, replaced having an entertaining evening. Similarly, if the ultimate end is expressed with sufficient generality, it may well be that only rather radical change would appropriately count as a change in the end: lesser alterations could be accommodated as changes in the way the ultimate end is specified rather than in the end itself.

introduced in connection with an ideally structured ultimate end.

The ideally structured ultimate end is so embodied in the corresponding agent's motivation that she would always actually decide, and view it as reasonable to decide, in its terms. For someone who is, perhaps, only on the way to having an ultimate end, and not an ideally structured one at that, that is too demanding a requirement. First, given that the end falls short of ideal structuring, there may not be an answer (or not one accessible to the agent) as to what acting in its terms is, and second, given that the agent may only be on the way to having the ultimate end, there may be slippage between the agent's actual motivations and those she would have at some later and more complete stage of the developmental process. What is reasonable to require is that the ultimate end or the developmental process be motivationally salient in approximately the sense in which Williams (1990) has claimed that reasons must be internal: there must either be some motivation to act in accord with the end, or, if there is no such actual motivation, there must at least be a sound deliberative passage from the agent's goals, preferences, dispositions, etc., to the acquisition of such motivation.

Similarly, it is too much to demand reflective stability of ultimate ends in some sense precluding reasonable alteration.[15] The perfect stability of the ideally structured ultimate end is only a function of the fact that, definitionally, it is provided against any changes in knowledge or situation, and so is never faced with anything radically

[15] Part of the reason is that it is less than clear what it would mean to preclude alteration when it is a developmental process – that is, a process of alteration – that is being supposed to stably preclude alteration. No doubt, something could be said along these lines as to the features or directions of change that must be included in the developmental process in order for it to count as unaltered, and change from which would, therefore, count as alteration, but I shall trouble neither to work out anything of the kind nor to attempt a showing that nothing of the kind is available, for there is a deeper problem, discussed in the text, with precluding alteration in ultimate ends.

surprising or wholly unanticipated. We, however, have no guarantee, definitional or otherwise, against the wholly unanticipated and, in particular, no guarantee against facing something wholly unanticipated in the light of which an ultimate end might reasonably be altered.

There are, however, several reasons for expecting a substantial measure of stability in ultimate ends. The most important may be that part of what is sought in an ultimate end *is* stability – for the ends by which one guides oneself to be mutually supporting rather than evanescent or interfering, for the ultimate end is something to guide one's life by, not just parts or episodes within it.[16] Other reasons are derived from the ultimacy and the generality of the end. Ultimacy limits the class of possible reasons for alteration, since there can be no other pursuits in terms of which to question or reject it – in contrast with all alterations of non-ultimate ends or objectives, which may occur in the light of other ends. The generality of the end insures that most changes will not count as changes of the end, but rather as improvements in the way it is specified or achieved.[17] So, though, on one hand, we cannot expect perfect stability, on the other, we must suppose that the ultimate end has *substantial* reflective stability. To go further than that rather general and imprecise claim, to address the question of exactly how much reflective stability is to be expected, is to go beyond the reach of any considerations that I can see to be available. That we expect substantial reflective stability is a general claim,

[16] The eudaemonist's conception of the ultimate end is, from the beginning, a conception of something that satisfactorily answers to concerns about living one's life, and not just parts of it, well. See § 4.2 and *NE* 1140a 25-28: "Now it is thought to be a mark of a man of practical wisdom to be able to deliberate well about what is good and expedient for himself, not in some particular respect, e.g. about what sorts of things conduce to health or strength, but about what sorts of things conduce to the good life in general."

[17] See note 15.

applying across many cases; what substantial reflective stability will amount to in particular cases will depend on the details of those cases.

In summary, to attribute ultimate ends (or the corresponding developmental processes) to real persons, the conditions that need to be satisfied include the person having reasons in terms of the end, the motivational salience of those reasons, and substantial reflective stability of the end. Where any of those features is absent, we have reason to question whether the person either has or is moving toward having an ultimate end. Where they are all present, there is an intelligible sense in which the ultimate end or the developmental process shapes the person's life.

5.3 Conflicting Ends: Problem and Solution

We enter the world with an initial motivational complement of biologically given ends. Some are evident very early, in the form of attempts to obtain nourishment, adequate warmth and comfort. Others, evident in such forms as desires for sex or status,[18] develop later.[19] Once they appear or develop, such ends may shape action throughout our lives.

The biologically given ends, however, provide only the starting point, for we are, to a significant degree, motivationally plastic. The ends by which we are motivated are not rigidly fixed. This fact has important consequences in three areas. First, we are

[18] That concern for status – or, more precisely, local relative status – has a biological basis is, I think, well-established. For explanation and some references, see Wright 1994, 236-262.

[19] In speaking of some ends as biologically given, I do not mean that they are immune to being affected by environmental vicissitudes. I mean only that the organism characteristically, and in normal environments, develops so as to have certain ends. They are biologically given in approximately the same sense as five-fingeredness is among humans: nearly universal in the species, in a way that can be explained in terms of the organism's genetic code, despite occasional exceptions.

subject to various processes of education and acculturation through which we come to acquire other ends.[20] We come to care about fairness or the prevention of suffering. We develop passions for chess or sports cars or philosophy. Second, as we mature, we become cooperators in our own motivational re-shaping. We deliberately cultivate tastes, acquire or strive to break habits, and more. Third, our motivational plasticity extends not just to making changes to or among the additions to the biologically given ends – to re-modeling the superstructure erected upon the biologically given foundation – but to reworking the foundation itself.[21] The importance of what is biologically given within the systems of our ends may be altered; it may even be set aside or over-ridden in the service of acquired ends. We take oaths of celibacy for a faith or become willing to die for a cause.

A further important feature, implicit in what has just been outlined, is that the ends by which we guide ourselves are *plural* and (to a significant degree) *mutually independent*. Even for the biologically given ends, and still more so for the acquired ends, there does not seem to be any way to represent them all as means to some over-arching end which is itself salient on the level of individual psychology.[22] In short, the

[20] I shall call the ends which are not biologically given *acquired ends*, though without intending any implication that the biologically given ends must be present or manifest from infancy in order not to count as acquired.

[21] A better metaphor may be Schmidtz's. He questions "the idea that starting points are what subsequently erected edifices must rest upon. We should not be fooled by the metaphor. We should realize that our starting points can be more like launching pads than like architectural foundations. A launching pad serves its purpose by being left behind." (1995, 76)

[22] The qualification is important because some might argue for an ultimate end not psychologically salient to the individual – at least not *as* an ultimate end – to which all others are means, e.g., reproductive success. I think that will not work, even on the level on which it is introduced, but for present purposes, it is sufficient to point out that reproductive success is not always what the individual aims at above all else. And if it is not *his* aim – something salient on the level of his individual psychology (rather than, in some metaphorical sense, the aim of his genes), it can neither be counted upon to solve his practical quandaries, nor even to provide him with guidance which he can recognize as relevant.

ends we have are plural and, to some degree, independent of one another. The independence of the ends is important, for that means that it is at least possible for one to be advanced at the expense of one or more others.

This kind of motivational complexity and plasticity is at the root of a problem which is virtually inevitable for us, but also provides the material for its resolution. In particular, what it provides is the material for an argument, grounded in instrumental reasoning, for the adoption of an ultimate end. I shall set out the argument briefly here and then elaborate upon its parts.[23]

(1) For a given person, there are conflicts among her ends.

(2) Given (1), it is impossible for all her ends to be achieved.

(3) Given (2), there is the problem, so long as the conflict persists, of inevitable frustration with respect to the achievement of at least some of her ends.

(4) The goal of finding a solution to the problem of inevitable frustration can anchor instrumental reasoning directed at eliminating or reducing conflict among her ends.

(5) An adequate solution will take the form of an ultimate end (or a developmental process directed towards coming to have an ultimate end) in terms of which the pursuit of multiple ends can be harmonized.

5.31 Why We Face Conflict Among Ends

So long as a person's ends are harmonious and realistically achievable, there may

[23] The following argument outline and section headings are not meant to precisely correspond to one another.

be no experienced need to evaluate the system of which they are part. With no difficulties, in principle, in guiding her actions by her ends, the practical task that faces her is just to find the means to those ends. That practical task, however, is unlikely to be ours. That is, it is unlikely for us to be so fortunate as never to have to deal with conflicting ends, to have an initial harmony among our ends. Our biologically given ends, combined with a perhaps haphazard overlay of acquired ends, almost inevitably lead to conflicts within the systems of our ends.

These conflicts are of two kinds. First, there is logical conflict, where, for the simplest case, one's end is both to bring about and to prevent the bringing about of some state of affairs, *S*. As discussed in Chapter Three, § 3.4, this is a real possibility, though unlikely for so simple a conflict. Second, there is circumstantial conflict, where there is no logical incompatibility between the ends, but where the circumstances are such that action that promotes one will tend to prevent the achievement of at least one other member of one's set of ends. Since the conflict in such a case is circumstantial rather than logical, it may be that there are options for taking action so as to alter the circumstances that give rise to the conflict. For simplicity, when I speak of circumstantial conflict, I shall assume that all such possibilities have either been exhausted or, in some other way, ruled out as unacceptable.

There are two reasons against expecting initial harmony to be a feature of our systems of ends. First, the ends that are biologically given were shaped by the evolutionary history that gave rise to them and presumably were adaptive – more so than available alternatives – at the time and in the environment in which they were shaped into their current form.[24] Even if it were plausible that under the conditions of the ancestral environment,[25] the biologically given ends would never come into conflict, that environment is not ours. What might have worked there cannot be expected to work in our different circumstances. Second, there is even less assurance that the acquired ends will not come into conflict, either with the biologically given ends or with each other. The calculational demands of insuring that the members of a large set of ends are consistent with one another are too great.[26] In the case of biologically given ends, we can say that part of the required "thinking" has been performed by natural selection; for the acquired ends, there is no one to do the thinking but ourselves – and, to avoid ever acquiring conflicting ends, much of that thinking would have to be carried out with immature cognitive capacities. Thus, it is not reasonable to expect there to be an initial or uncontrived harmony in a person's set of ends. An harmonious and realistic set of ends may, for us, be an achievement, but it is not a starting point, not an initial harmony, that is simply given or to be taken for granted.

5.32 Conflict of Ends as a Problem

Conflict of ends is a problem, but before proceeding with its discussion, there is a doubt to be addressed: Why is the conflict to be described as a *problem* – something that

[24] See Chapter One, especially § 1.23.

[25] I think it is not plausible, because (a) evolutionary processes can only be expected to improve upon existing alternatives, not to achieve perfect adaptation to an environment, and (b) it is something of a misnomer to speak of *the* ancestral environment with respect to the evolution of biologically given ends, since they are almost certainly the result of accretion and refinement from *many different* ancestral environments.

[26] Consistency-testing is subject to combinatorial explosion. The more items that must be tested for mutual consistency, the harder the problem is – and the difficulty increases disproportionately as the number of items to be tested increases. See Chapter Three, § 3.4.

stands in need of a response or solution – rather than as a condition that must simply be accepted? Part of the answer is that the force of the “must” in “must simply be accepted” presupposes that nothing can be done to change matters in any relevant respect, and also that “simply accepting” is not itself something that can be carried out in different ways that can be distinguished as being better or worse responses, as, for example, acceptance with good or with ill grace. Perhaps it is true that no response is better or worse than any other, but, if so, that stands in need of further support. It is no more to be taken for granted that nothing can be done than that something can. Another part of the answer is that conflict of ends, if nothing is done about it, insures frustration; action on behalf of one or some ends will guarantee the failure to achieve one or some others. That frustration in the pursuit of one’s ends is a problem in the sense of calling for some response I take to be very nearly analytic: it would be difficult to understand how some alleged objective of a person really was his *end* if he regarded his frustration in achieving it as entirely unproblematic.[27]

To return to the main line of discussion, where only a single end or a harmonious set of ends is in question, we can manage with notions of relative efficacy. But by late adolescence or early adulthood, if not earlier, we become reflectively aware of conflict among our ends, that the pursuit of one requires the frustration of another. It is uncontroversial that means can be graded as better or worse relative to a given end. What is not obvious is how to grade actions undertaken as means when the ends themselves seem to point in different directions.[28] When there are conflicts among ends, what is

[27] Perhaps, he might *find* it impossible to do anything about the conflict of ends. Then, the frustration might be regarded as *tragic* rather than problematic.

[28] This problem, I think, is one of the things at the root of our desire to grade lives or systems of

better in terms of one will be worse in terms of another, and without begging the question in favor of some end or subset among our ends, there will be no answer as to what course of action best serves our ends. In the absence of an answer, we face the problem of what to do about the conflict.

In fact, there are two problems here, the *local* and the *global* problems of conflicting ends. There is the local problem of what to do in the particular instance, and there is the global problem of what to do about the general fact of conflict among our ends.

In the particular instance, it must be decided whether to direct our action to the service of one end or the other, or perhaps neither, of a conflicting pair, when both cannot be pursued.[29] In the absence of a reason to go one way or the other, one may be selected arbitrarily, or some other *ad hoc* procedure may be applied. Or the local problem may be addressed by way of an attempt to resolve the global problem.

For the local problem, so long as there remains for the agent a pair (at least) of ends in conflict, no fully satisfactory solution is possible. Whatever is done, since action which serves one end will disserve the other, will amount to acting against at least one and possibly both of the conflicting ends. The only thing that could count as a fully satisfactory solution to the local problem would be something that removed the conflict. Since, whenever there is conflict between ends, the ends must be independent with respect to one another, the minimum condition for removal of the conflict is either to

ends as wholes as being better or worse. When ends are harmonious, we can just say ‘better for this end’ or ‘better for that other end.’ It is in the face of conflict among our ends that we raise the question of how to grade this end and that, either in comparison to one another or as parts of larger systems of ends.

[29] The case of two ends in conflict is, of course, only the simplest form the problem takes. More elaborate conflicts are also possible.

remove one of the ends or else to remove their independence.

In other words, a possible satisfactory solution would be to abandon at least one of the ends in question, where abandonment would imply more than just non-pursuit, but rather abandoning it as an object of pursuit – which would mean that it is no longer sought or aimed at for its own sake, and therefore no longer in conflict with anything else sought or aimed at for its own sake. Another way would be to establish some kind of priority relation, suitable to arbitrate conflicts, between the conflicting ends. Either one of these would mean that at least one of the ends would cease to have its former status. In the former case, this is obvious, for at least one of the ends ceases to be an end. In the latter case, both could remain ends, but the priority relation itself, or something from which it derived, would have to have the status of an end,[30] to which the formerly independent and conflicting ends would serve as means.

In principle, local problems could be addressed piecemeal as they arise, with the abandonment of some end or the establishing of a priority relation whenever a conflict is discovered. But that seems inadequate in more than one way.

First, the particular changes introduced into the system of ends, if they are only directed piecemeal to resolving conflicts as they arise, will lack any rationale beyond the fact that they do indeed serve to resolve the particular conflict.[31] The question to be faced

[30] Suppose a person has a pair of mutually unranked ends, E_1 and E_2, that upon occasion come into conflict with one another, and that, to resolve the conflict, she adopts the simplest possible priority relation between them, one which, say, selects E_1 for pursuit in any case of conflict with E_2. Then, it might be said that there is no *new* end, just the same ones with a new ordering. I take this difference to be purely terminological, and to turn upon how one individuates ends. The fact remains that either the priority relation does something that the ends alone did not do, or else at least one of the ends has changed in weight or importance in decision-making. If one end is now more important than the other, whereas before it was not, and if ends are individuated by differences in the courses of action they license, then the ends are not the same as before the adoption of the priority relation.

[31] Suppose there were some further rationale for resolving a conflict one way rather than another.

is: Why resolve the conflict in *that* way? Why eliminate the end, E_1, rather than E_2? Why adopt the priority relation, R_1, rather than one of its alternatives, $R_2, R_3, ..., R_n$? The answer might be just that the selections made *do* resolve the particular conflicts, though something else would have done so as well. That would be to admit that there is no further rationale beyond their role in resolving the immediate conflict. In effect, the argument for selecting, say, E_1 rather than E_2 will be that *something* is needed to resolve the conflict, and this is something. Only if nothing better could be provided would that be rationale enough.

Second, so long as the approach to conflicts of ends is piecemeal, the changes introduced may not reduce occasions for conflict. In particular, though an end-elimination will always reduce somewhat the possibilities for conflict between ends, since there will be one less end to come into conflict with any others, the adoption of a priority relation, since it involves adopting some new end, may well create additional occasions for conflict.

Third, the attempt to address conflicts within one's system of ends solely through piecemeal adjustments is psychologically unrealistic. Abandoning or acquiring an end is not the work of a moment. An end is something that has a more or less enduring place in one's motivations. It is not as if one could simply press a button and instantly have an altered set of motivations. (If there were a costless way to instantly alter motivations, why wouldn't everyone be ecstatically happy, since they would only need to alter their motivations to be utterly delighted with their activities and circumstances?) Given,

Then, to that extent, the conflict would not be being addressed in an entirely piecemeal way, for the rationale would be generalizable and therefore applicable to other conflicts.

however, that ends and the corresponding motivations are not instantly altered, that eliminating one end or acquiring another takes work, what is going to keep the person at the task? Since we are supposing that the adjustment is piecemeal, it won't be that he sees a reason for eliminating the end, E_1, rather than E_2, or for adopting the priority relation, R_1, rather than R_2. In the absence of a reason and in the face of contrary motivation – to keep on pursuing the end selected for elimination or to ignore the new priority relation – it is likely that particular piecemeal adjustments will not be successfully completed and even more likely that all the piecemeal adjustments needed to remove conflict among one's ends will not be successfully completed.

Last, piecemeal adjustments fail to get to the root of the matter. At best, they remove local problems as they arise or are noticed, but do nothing about either the fact that the person's system of ends is such that conflicts do arise or about the fact that new ends adopted, whether to resolve earlier conflicts or on some other basis, are also apt to engender conflict. What is not addressed by piecemeal adjustments is the global problem: Is there something to be done about the general fact of conflict among ends, which gives rise to the various local problems and their particular frustrations?

If anything can be done about the general fact of conflict, something that eliminates or reduces it, that will be, *ceteris paribus*, a better solution to the global problem than piecemeal adjustments in response to local problems.

5.33 Removal of Conflict as a Maieutic Objective

At some point in our lives, we face the global problem of conflicting ends, which consists in the general fact of conflicts within our systems of ends. That general fact has two major features, first, that, for each of us, our systems of ends include conflicts which are discovered from time to time, and second, that the processes by which our systems of ends are altered, including both the piecemeal adjustments aimed at resolving already discovered conflicts and other forms of the acquisition of new ends, are apt to introduce new conflicts. The problem, then, has both synchronic and diachronic dimensions. There are the conflicts existing at a time, and there are processes of end-alteration which themselves may give rise to further conflict. An adequate response will need to address both dimensions.

We can let solving the global problem stand as a specification of a goal to which instrumental reasoning is anchored, and ask what will serve to solve that problem. In principle, though on a broader scale, the same possibilities are available as for solving various local problems: Since the problem consists of conflict of mutually independent ends, plus, of course, the various processes that lead to further conflict, the solution will have to either eliminate ends or eliminate their independence, and will, in addition, have to provide some way of regulating or channeling the acquisition of further ends so that they are less liable to give rise to conflict.

There are two basic possibilities for the solution to the global problem. Either it will involve an ultimate end, whether in the form of acquiring one or of undergoing the corresponding developmental process, or it will not. The latter possibility can in turn be subdivided into approaches which involve acquiring some additional end and those which do not.

The last of these, an approach to the global problem that involves no end-acquisition, is not a real option, for there is only one way to remove conflict between ends without acquiring some new end. That is to eliminate one or more of the conflicting ends. The problem is in finding a rationale for end-elimination, for abandoning one end in particular rather than another. There must *be* such a rationale, or we would only be engaged in the kind of piecemeal adjustment already dismissed as inadequate. But any such rationale will either be entirely *ad hoc*, and thus only verbally distinct from piecemeal adjustments, or it will be generalizable. Even a rationale so simple as conflict-avoidance (as distinct from this-conflict-avoidance) will apply to other conflicts and will imply that some end-eliminations are better than others.

For example, it may be that there is a conflict between ends *A* and *B*, between *C* and *D*, and between *B* and *D*. Examined one at a time, *A* might be eliminated from the *A*-*B* pair and *C* from the *C-D* pair, leaving *B* and *D* in conflict. If there is no other ground than the elimination of conflict, the better option would be to eliminate *B* and *D*, and thus, at the cost of abandoning two ends, to eliminate three conflicts. Even to go that far is to apply *some* generalizable rationale to multiple cases of conflict. Counting conflicts and settling upon which ends to eliminate on the basis of reducing the total number of conflicting end-pairs may of course be too crude, and it is easy to imagine more refined criteria, depending upon the details of the case, but more refined criteria would *also* be generalizable rationales.

To the extent that generalizable rationales enter the picture, even the simple elimination of a member of a set of conflicting ends serves to reduce the mutual independence of remaining ends, even if they are not themselves members of the same conflicting set, for if, out of the set of conflicting ends containing *A* and *B*, *A* is eliminated because, in addition to being in conflict with *B*, it has the property, *F*, then, in any other conflict, if one and only one of the ends in conflict, *C,* also has the property, *F*, then *C* will be the one to be eliminated, unless it also possesses some offsetting property, *G*. In other words, what is to be done to resolve the *C-D* conflict will depend in part upon the resolution to the *A-B* conflict, or, more precisely, upon the reason for the choice that was made there.

Thus, if end-elimination is to be more than piecemeal adjustment, we must appeal to generalizable rationales. What is their status? It appears that they must either be ends or else somehow derivative from or based upon an end or ends other than those immediately involved in the end-elimination at hand, and further, that the rationale or the ends upon which it is based are regarded as more important than at least one of the ends involved in the conflict. For if other ends were not involved, or if any other ends involved were not regarded as more important than the end to be eliminated, it would be unclear why something aimed at for its own sake should be given up to comply with the rationale.

The upshot of the argument so far is that all approaches to the global problem involve the acquisition of or appeal to new ends.[32] Thus, the objective of solving the global problem of conflict is a maieutic objective, one which can be achieved only by coming to have an end or ends.

And given that some ends must be acquired if the global problem is to be approached at all, it is straightforward that there is no stable stopping point short of an ultimate end, for the alternative is to continue to acquire ends in the form of priority relations or generalizable rationales, which themselves may engender further conflict. To

[32] See note 31.

be sure, it *might* be that the successive acquisition of ends to address conflicts as they arise would lead in the direction of progressive simplification and unification of the system of ends. But if so, that would be accidental, for it might be that conflicts between *A* and *B* are addressed by a priority relation, R_1, that conflicts between *C* and *D* are addressed by a different priority relation, R_2, and that, with two new ends in the picture, there are conflicts between R_1 and R_2, between R_1, *C* and *D*, and between R_2, *A* and *B*. Three conflicts may be substituted for two. If there is some feature of the process of end-acquisition addressed to conflict-removal that prevents that kind of outcome – that insures that, over time, the system of ends becomes (or tends to become) progressively simplified and unified, that would be equivalent to having, or moving in the direction of having, an ultimate end.

In summary, given the plurality and mutual independence of our ends, there is the problem of almost inevitable conflict, and, when there is conflict, the achievement of some ends insures the frustration of others. Given the conflict, the search for a solution is a maieutic objective, which can in the end only be achieved by the acquisition of a new end. More specifically, given conflict and end-plasticity, there may be a solution. By revising or adjusting the set of one's ends – perhaps acquiring new ends, perhaps eliminating some, perhaps altering relative weights – one can reduce or perhaps eliminate end-conflict and the attendant frustration. In effect, this amounts to adopting an over-arching end to which the formerly independent ends become constitutive means. The over-arching end prescribes something that, without a fair amount of background (which I have tried to provide), might sound nearly empty, namely, successful end-pursuit or, perhaps better, since it is more obviously related to eudaemonism, *comprehensively* *successful living*, where a life can be said to be comprehensively successful to the extent that it is a success in all the ways that a life can reasonably be expected to be a success.

5.4 Structuring the Ultimate End

We can give instrumental reasons for adopting or moving in the direction of having an ultimate end, but, even when spelled out as comprehensively successful living, there is more that we need to know about the component ends – about what one needs to succeed in doing in being comprehensively successful. Can anything general be said about the component ends? If an ultimate end gives the shape of a life, what is the shape of a life guided by comprehensively successful living?

Much of the answer will be subject to individual variation. Consider what may be called *endowments*, which include both physical and mental capacities and potentials and access to external resources. Given a set of endowments, some aims or projects will make sense, while others will not. Thus, for example, paraplegics, not to mention most of the rest of us, are unlikely to succeed in plans requiring exceptional athletic skills, nor are the very poor likely to make their livings as investment bankers. Since people differ in endowments, they will also differ in the kinds of aims or goals that it will make sense for them to pursue.

Further sources of variation will be traceable to features of the individual's motivations. These might also be classified as endowments, but they are significant enough to merit separate mention. These may differ from one person to another, not only through differences in environment and experience, but also by way of innate predispositions brought to experience.[33]

On the plausible assumption that these sources of variation between persons will not somehow be abolished or counteracted through the developmental process involved in the acquisition of an ultimate end, what will count as successful living will depend both upon what one has to work with in the form of endowments and upon the innate and acquired motivational features in terms of which one judges what to do with one's endowments. What these considerations amount to is that, so far as the sources of variation considered have an impact upon ultimate ends and, through them, upon the lives shaped by those ends, there is not a general answer to questions about the component ends involved in comprehensively successful living.

But there must be commonalities as well. It is hardly adequate for an ethical theory to say that everything depends on the individual. So, given that there will be much that properly varies between persons, depending upon their endowments and aims or projects, are there any features that we can argue that they nevertheless should have in common?

What I shall argue is that there are reasons for accepting and coming to embody in one's motivations and behavior practical principles having the functional role of the virtues. Traits fulfilling this role, I shall call *f-virtues*. F-virtues have at least the

[33] Impressive evidence of the innateness of some psychological features can be found in the studies of identical twins, separated at birth, cited by Pinker. "Their minds are astonishingly alike, and not just in gross measures like IQ and personality traits.... They are alike in talents such as spelling and mathematics, in opinions on questions such as apartheid, the death penalty, and working mothers, and in their career choices, hobbies, vices, religious commitments, and tastes in dating. Identical twins are far more alike than fraternal twins, who share only half their genetic recipes, and most strikingly, they are almost as alike when they are reared apart as when they are reared together." As he concludes, "by showing how many ways the mind can differ in its innate structure, the discoveries open our eyes to how much structure the mind must have." (1997, 20-21)

following features[34]:

- They are stable traits of character which, in appropriate situations, issue in action.
- They involve intelligent responsiveness to relevant features of those situations.
- They are partially constitutive of the ultimate end, which, for present purposes, I am identifying with comprehensively successful living.
- As constitutive of the ultimate end, they are ends and therefore cultivated and exercised for their own sakes.
- As constituents of the ultimate end, they constrain, though they do not dictate, what else may be part of the ultimate end.[35]

The reference to their functional role is deliberate, despite the fact that the best examples of f-virtues are also, simply, virtues. The point is to focus upon these functional features first without demanding that whatever possesses these features also be among the traditionally recognized excellences of character.

That there are stable traits of character that involve intelligent responsiveness, on cognitive, affective and motivational levels, to situations of various types – greed,

[34] These features are discussed at greater length in Chapter Four, § 4.4.

[35] The fifth item is not so much an additional requirement as an entailment. Any trait of character that satisfied the other conditions would also constrain what else could be part of the ultimate end.

gentleness, generosity and fairness, for example – I take for granted.[36] In light of that, remember that the upshot of earlier argument was that an adequate solution to the problem of conflicting ends would take the form of an ultimate end, specified for our purposes as comprehensively successful living, in terms of which the pursuit of multiple ends can be harmonized, and which is constituted by those same ends.[37] If any of the ends constitutive of comprehensively successful living takes the form of an action-guiding trait of character, then it would be an f-virtue, and if an instrumental case can be made in favor of including one or more f-virtues among one's ends, that would complete the project of showing that instrumental reasoning yields the eudaemonist structure.

What instrumental reason can there be for an agent to cultivate an action-guiding character trait and practice accordingly? The answer depends on several factors. First, there must be a recurrent situation-type for the character trait to be responsive to and exhibited in. Second, the trait must be one which can be built up and established as part of the agent's character through learning and practice. Third, possession of the trait must be advantageous in some way, as assessed from the agent's standpoint.[38] Fourth, the trait must facilitate decision-making, action and response appropriate to the relevant situation-type. Perhaps the most important, though not the only, type of facilitation here consists of circumventing the need for calculation when time is short. When one already has a settled disposition to respond in a certain way, there is less need to figure out what to do.

[36] For some doubts, however, see Harman 2000 and 2003. For some discussion, see Flanagan 1991, 276-314, and for briefer discussion with a response, see Flanagan 2002, 153-159.

[37] See §§ 5.3 - 5.33.

[38] This locution is not meant to be so narrow as "advantageous for (or to) the agent." The agent may assess a trait as advantageous because of its contribution to something for which he cares, which may or may not be some advantage to himself.

Fifth, there must be some reason that the trait stands in *need* of cultivation, that it is not something for which one can simply count upon uncultivated tendencies. The virtues – and therefore also their functional equivalents, the f-virtues – are, as Philippa Foot says, "*corrective*, each one standing at a point at which there is some temptation to be resisted or deficiency of motivation to be made good." (Foot 1978, 8) In a similar vein, Walter Lippmann comments that

> [t]hey would not be called virtues and held in high esteem if there were no difficulty about them. There are innumerable dispositions which are essential to living that no one takes the trouble to praise. Thus, it is not accounted a virtue if a man eats when he is hungry or goes to bed when he is ill. He can be depended upon to take care of his immediate wants. It is only those actions which he cannot be depended upon to do, and yet are highly desirable, that men call virtuous. (1957, 207)

Taken together, these features explain why an agent would have an instrumental rationale for cultivating f-virtues. The first and second features insure that the trait can be acquired and that there are circumstances apt for its acquisition and exercise. The third provides the reason for acquiring it – its advantageousness. The fourth and fifth together explain why the f-virtue has to be cultivated and why the corresponding activity and response must be practiced or engaged in for its own sake[39] – and therefore as a

[39] More precisely, the f-virtue will initially be cultivated for the sake of something else, its advantageousness, but *what* is cultivated is the disposition to respond and act in certain ways for their own sakes, not just for the sake of something else. See above, § 5.1.

constituent of the ultimate end[40] – for otherwise, the advantageous response cannot be counted upon to be forthcoming.

5.41 The Advantages of Virtue

The question that remains is whether the five conditions are satisfied and, therefore, whether an instrumental case can be made for any f-virtues. Since the best argument I know that the conditions are met in fact takes the form of arguing for what I have been calling the traditionally recognized excellences of character, that is what I will present.

There is a further advantage in directing attention to the members of the standard catalogue of virtues in the fact that it is uncontroversial that they satisfy four of the five conditions. Specifically, it is uncontroversial that there are recurrent situation-types for the virtues to be responsive to, that the virtues can be established in one's character through learning and practice, that they facilitate decision-making, and that they are corrective. What remains is to argue that they are advantageous. What I shall do is sketch, but no more than sketch, an argument for the advantages, from the agent's perspective, of the traditionally recognized excellences of character, the virtues.[41] Since

[40] If an f-virtue is practiced for its own sake, but not as a constituent of the ultimate end, then it must be as a means to some other end. But, for the kind of dispositional trait under consideration here, that will often be implausible, for a situation-type to which an f-virtue is responsive will not be confined to the pursuit of some single end to the exclusion of others. Courage, for example, involves a kind of response to dangers of all types and faced in the service of many different ends. If courage really is an advantageous disposition to have, it will not be confined to being a response to danger of one or a few types, depending upon what end is being served.

[41] Part of the reason for the sketchiness is that, for present purposes, I take the traditionally recognized excellences of character in the aggregate, without considering separately the claims or merits of honesty, courage, compassion, fairness and so on. Nor am I addressing the difficult problems associated with the fact that some character traits once, and perhaps traditionally, regarded as virtuous – chastity, for instance – may seem less compelling now. For pertinent discussion, Martha Nussbaum's " Non-relative

the benefits or advantages are to be taken into account as reasons for living a virtuous life, they have to be in a form that can be understood by the agent, prior to her acquisition and practice of the virtues. Any advantages of the moral or virtuous life that can only be appreciated from within will not belong in the instrumental case. The boundary between what can be appreciated from outside or from within, however, is not sharp, for what can be appreciated from outside the virtuous life may include acknowledgement of facts that can be fully appreciated only from within. That virtue is its own reward may only be fully understood by the virtuous, but that does not mean that the outsider cannot see that the virtuous *do* find satisfaction in virtuous activity, not just in external goods to which it leads, and therefore does not mean that the outsider cannot see that there is some reason to suppose that, were she to become virtuous, she, too, would find satisfaction in the virtuous life.

Though they are real considerations, such appeals to the intrinsic rewards of virtue are not the main part of the case. What is more important is the fact that what are widely recognized as virtuous traits of character have a systematic tendency to be advantageous to their possessor.

Note first that traits that are thought to be generally advantageous neither to the possessor nor to others will not be recognized as virtues. At best, such traits will be thought matters for indifference, and if they are actually generally disadvantageous to either the agent or to others, they will be regarded as failings, vices or perhaps simply misfortunes. That will include, of course, any traits that are generally advantageous to the agent but disadvantageous to others. It will be especially important for a society to

virtues: an Aristotelian approach" (1993) is quite interesting.

discourage the development of such traits.

Since we can count on any traits that are socially recognized to be generally harmful to others to be discouraged, only traits that are either advantageous to the possessor, to others or to both will be recognized as virtuous.

That means that two out of the three possible combinations of individual and social advantage – the case in which the agent gains (without disadvantaging others) and the case in which both the individual and others gain – include advantage to the agent. The troublesome case is that in which there is some socially advantageous trait which is not advantageous to the individual. This may take either of two forms – the easier case in which the trait is merely not advantageous to the agent without being disadvantageous and the more difficult in which the trait appears actually to be disadvantageous to its possessor. The question for the troublesome case is how such traits are elicited, and the answer appears straightforward: the development or possession of the traits is rewarded in various ways, with praise, honor or respect, as well as with the more tangible and often, though indirectly, associated rewards in such forms as wealth and influence.[42] What these facts mean is that the virtues, even in the troublesome case in which performance seems sometimes at odds with the advantage of the agent, tend to be advantageous to the agent. At least, it tends to be advantageous to the agent to acquire the virtues, though acting accordingly may of course be disadvantageous in the particular instance. This should not be surprising. As R. M. Hare notes:

[42] The indirectness of the more tangible rewards, combined with the fact that they are not guaranteed to materialize, may be essential to insuring the real development of the virtuous traits as opposed to their simulation. Direct, tangible and relatively assured reward might make it psychologically impossible to develop the authentically virtuous trait, an aspect of which is performance of the relevant activities for their own sakes.

> It is a physical and not a social fact that there are no rings of Gyges. But the more important empirical facts here are social ones. It is no accident that the world and society are such that crime does not in general pay. People have made it like that because they did not want crime to pay; it is more in the general interest if criminals are brought to book. We must not think here merely of the legal system and courts and policemen; they ... would be ineffective unless backed up by much more powerful social pressures. Mankind has found it possible to make life a great deal more tolerable by bringing it about that on the whole morality pays. It is better for nearly all of us if social rewards and penalties are attached to socially beneficial and harmful acts; and so it has come about that on the whole they are.[43]

Much more could be said along these lines, and further and deeper investigation would surely be desirable. But what has been said so far, I think, provides a reasonable case that the virtues are systematically advantageous to their possessor. Though it is true that virtue may require significant sacrifice, it is an illusion, perhaps due to misplaced emphasis, that makes it appear that the virtuous life is dominated by sacrifice and cannot be expected to be good for the virtuous. The truth is more nearly the reverse: the advantages are the dominating feature and the sacrifices the occasional exceptions. Given

[43] Hare 1981, 195-196. Also of considerable interest in the current connection is the entire chapter from which the quote is taken, 188-205.

this – and given the fact that the advantages appealed to can be appreciated by an outsider who is not assumed already to be virtuous – the instrumental case for incorporating the virtues as constituents of comprehensively successful living appears to be in good shape. We are well advised to make virtue a part of our lives.

5.42 The Maximizer's Challenge

The eudaemonist structure, in which there is an ultimate end of living well, which is partially constituted by commitment to the practice of the virtues, is, when embodied in our motivations, actions and responses, good for us. It might be replied that more is needed from a credible instrumental argument. If we are to be instrumentally justified in undertaking to acquire the virtues, we need assurance that it is the best option. And that, of course, is something I have not provided, nor am I in a position to provide it. There are, however, several levels of reply available.

Before beginning to reply, I shall distinguish two sorts of concern that may be expressed by the objection. One concern is with the virtues in general, with whether it can be good to adopt and so internalize practical principles that it is psychologically difficult or impossible to violate them to secure great benefit or to escape great harm. Wouldn't one do better to be more loosely attached to one's principles? The other concern is whether the traditionally recognized excellences of character are the right principles to internalize, whether there might be some other f-virtues with which one would do better. The two are only partially independent, and to the extent that they are not, the same considerations apply to each, but there are some differences as well.

The first concern is in the same spirit as the familiar act-utilitarian criticism of rule or indirect utilitarianism. This critic will prefer rules of thumb, generalizations to which one expects exceptions, over any principle or character trait that cannot, without difficulty, be violated when it is advantageous to do so. To have a convenient abbreviation (at some sacrifice in accuracy), let us call the kind of principle to which the critic objects a *non-violable principle.* The question the critic asks is: if some non-violable principle is adopted, supposedly upon instrumental grounds of advantage to the agent, why is not the possibility of great disadvantage attendant upon following the principle a reason to give it up, or, more precisely, to give up or never adopt in the first place the non-violability feature? Why, in other words, is not a *violable* principle better?

There are two points to be made in response. First, the critic can be asked if he makes it a non-violable principle (a non-violable meta-principle?) to avoid adopting non-violable principles. I wouldn't attach much importance to the charge that his position logically undercuts or refutes itself, but there is a serious question whether it might be advantageous sometimes to adopt a non-violable principle rather than a violable principle. That appears to be an empirical question, not the sort of thing to be settled from one's armchair.

Second, it can be granted that we might do better if we were prepared to violate our principles just when it would be to our advantage to do so. Adopting a policy, however, of guiding oneself only by violable principles is only sensible to the extent that we are reliable judges of when it would be advantageous to violate them. The fundamental problem with this is that it rests upon a tacit assumption that we have unlimited cognitive and motivational flexibility at our disposal. If our capacities are limited, as they surely are, and if in particular we are unlikely to make the best decision under pressure – perhaps due to bias or some other vice, it may well be that we would have done better to adopt some non-violable principles than to insist on guiding ourselves only by violable principles.

The second type of concern, whether the traditionally recognized virtues are the best principles to internalize, poses a different sort of problem. This second critic is not concerned so much that internalizing any principle, making of it an f-virtue, will tie his hands when it would be advantageous not to have his hands tied. Rather, his concern is that there may be better principles to internalize than the traditionally recognized virtues.

I agree immediately that I have no proof that the traditional virtues cannot be improved upon. Nevertheless, several things can be said in their favor. The first is that it is not a terribly weighty consideration against the virtues so long as no particular alternative is proposed. Once particular alternatives are proposed, some non-traditional f-virtues, then they can be subjected to examination and compared with their more traditional rivals.[44] Until then, the challenge is only theoretical.

Second, the widespread recognition of the virtues is evidence that they have proven satisfactory over a broad range of experience and for long periods of time. In every domain of inquiry, we must start where we are, and if we find ourselves with moral beliefs – in this connection, especially beliefs about the virtues which are largely the deposit of our education and socialization – there is no reason to reject those beliefs simply because we lack proofs that they are right.[45]

[44] If the virtues are most fundamentally characterized in terms of appropriate responsiveness to situation-types (see Nussbaum 1993), then the imagined comparison may be between competing specifications of the same virtue.

[45] In the unlikely event that someone approaches the question without having any beliefs about what is virtuous and what is not, I would urge, among other things, the evidential value of widespread

These answers are at best only partial, however. The charge that I have not shown that it is best to acquire and practice the virtues is one that, so far as I can see, cannot be met in that form. There is, however, an underlying assumption behind that charge that deserves to be challenged in its turn. This is the assumption that a satisfactory instrumental case for virtue should be a case that exhibits virtue, or perhaps acquiring the virtues, as maximizing – as doing the best that one can, given whatever constraints are relevant.

That assumption, however, is one that I criticized at length in Chapter Two.[46] We are unable, especially in connection with problems having the largest scope, to maximize. Our preferences are not fully ordered, and among options not fully ordered by preferences, maximizing has no determinate reference.

Since we cannot reasonably expect the choice to acquire and practice the virtues to be a maximizing choice in any event, the fact that no such argument has been provided is not a failing in the case for the virtues. In the place of maximizing, something else must be substituted. We must select, not what is best, but something that is good enough. It is true that no argument has shown that there are no available improvements upon the virtues. It is also true that no argument has shown that some better possibility will not come to light tomorrow or next week. Those facts do not detract from the case that we have instrumental reasons to acquire and embody the virtues, for they are advantageous to the agent. What I have offered aims, not to be a proof that it is best to be virtuous – if I

recognition. Also of interest is the fact that some evidence suggests that we are carriers of evolved moral predispositions. See, for example, Pinker 2002, Chapters 11 and 15, and Wilson 1993. If so, the limits to what we can find satisfying may be narrower than we suppose, and the evidence of traditional recognition becomes more powerful.

[46] See especially §§ 2.31 – 2.34.

am right, no such proofs are available – but to be good enough.

5.5 Directions for Exploration

There are questions which bear upon directions for further exploration and development which I have not been able to address here. About some of these, I shall try to indicate what the questions are and why they remain problematic.

5.51 The Instrumental Framework: Costs and Benefits

One of the most obvious questions arises from the fact that I have attempted to operate within an instrumental framework: How far can an instrumental approach in ethics be expected to go? The short answer, I think, is that it has significant reach – more than many have supposed – but that it is still limited. One limit is not so much to what can be expected of an instrumental approach, but to what has been offered here. The case for a form of eudaemonism incorporating the traditional virtues has been sketchy. I think much more can be done in the way of filling in the details, and is worth doing, but that must remain for another time.

Setting aside the sketchiness of the argument, an important limit is that, to the extent that the case for eudaemonism, and, in particular, for acquiring and practicing the virtues, is instrumental, its cogency will vary among addressees. The basic reason is that an instrumental case for doing anything can be represented as comparing expected costs and benefits. It is unrealistic to suppose that they will balance in exactly the same way for everyone, and even if, implausibly, the instrumental case could promise gains to everyone, the gains might be insufficient to make the costs to be undergone worth the bearing. Even if there is always a benefit, there may not always be a net benefit.

A particularly important application is related to the fact that much of the cost must be borne at the beginning, in the form of effort, practice and habituation to acquire the virtues, while much of the prospective benefit, for the sake of which the cost is borne, is to be found in the form of ongoing returns in the more distant future. Since our lives are limited, the later one gets underway, the less is the chance that the future returns really will justify the costs borne. And it is not just that later in life, there is less time to recover the costs: The costs themselves increase, as habits, dispositions and value-judgments become more firmly a part of one's character, and therefore, more difficult to alter or excise, should it be necessary.

This fact about the timing and amounts of the costs and benefits suggests that, for most people, there may come a time at which no satisfactory instrumental case for the virtuous life can be made, for there will not be a sufficient future in which to recoup the costs. Accordingly, if there is an instrumental case for the virtuous life at all, it is also a case for getting started early. That in turn suggests that it is important to begin inculcating virtue at an early age. In other words, moral education, beginning early, is important. So far as we want people to develop virtuous characters, it is unwise to unnecessarily delay the task of leading them to do so.[47]

A further limitation is that the instrumental case depends upon assumptions about normal motivations. What, if anything, can be said about or to people with atypical motivational repertoires? It is hard to say anything generally applicable to such cases, but part of the answer will depend upon how atypical the motivational repertoires in question

[47] There is the further point that, though the instrumental case for someone to become virtuous is framed in terms of benefits from the agent's perspective, there are also benefits to the rest of us in dealing with and living among virtuous people. From *our* perspectives as well, there are reasons to encourage the development and practice of the virtues as early as feasible.

are, and in what way they are atypical. However, we cannot count upon there being a satisfactory instrumental case that everyone, regardless of motivations, has reason enough to acquire and practice the virtues. An argument meeting that condition would be nice to have, but I cannot see that it is available.

5.52 Beyond an Instrumental Approach?

Since both of these are limitations upon the reach of an instrumental case, they suggest the further question whether the second thesis of instrumentalism, as earlier defined,[48] is true. Is practical reason in fact confined to regimenting actions understood as means to the ends that they serve? I have argued in effect that the restriction of the scope of practical reasoning to the service of means to ends, even if true, can go further than many imagine. A question calling out for exploration is whether the thesis *is* true: is there something that can be offered to those who are (rationally) unmoved by the instrumental case,[49] something that can legitimately claim to be reason rather than bludgeoning or propaganda?

I am inclined to think or to hope that the answer is affirmative, and am interested in exploration along broadly Kantian lines. What many philosophers inspired by Kant have sought has been a grounding of all of morality in reason alone, with no need to appeal to any sentiment or commitment that could be otherwise. If that kind of grounding can be provided, it would reach to and have a grip upon every rational being. There would be no barriers constituted by unusual motivations or lack of time for prospective

[48] See § 5.0.

[49] That someone may, in some other way than rationally, be unmoved by good reasons is, of course, not a disease for which philosophy offers any remedy.

benefits to materialize. The case for morality would be rationally compelling for all. I think that inspiring and that there are real prospects of fruitful non-instrumental approaches, but even if it is true that practical reason is not confined to instrumental rationality, it may not go so far as Kantians would hope.

5.53 The Politics of Virtue

The kind of eudaemonism or virtue ethics I have favored is most naturally understood as providing guidance, primarily or in the first instance, for individuals. The questions to which it is addressed are, in decreasing order of generality: What is it to live well? What is it for *this* person to live well? What is it to act and respond well in *this* situation? But there are other questions to which a eudaemonistic approach does not so easily lend itself, and which have thus been underexplored.[50] These are, broadly speaking, *political* questions: How is a society to be ordered? What can appropriately be required of everyone?

On a general level, it is possible to say what should be done about political issues. The same generic advice as applies to other contexts and types of problems can be given here: The right thing to do in a given situation – including, presumably, a given political situation – is to act as the person of practical wisdom would act. The principal problem is lack of specificity: What *would* a practically wise or virtuous person do with respect to

[50] Part of the problem is not that the politics of virtue has gone unexplored but that the exploration has been undertaken by thinkers, primarily the ancient Greeks, who were facing so different a political world from ours that it is difficult to apply lessons, in other than the most general terms, from their exploration to our situation. Michael Slote, with somewhat different concerns than mine, discusses the issue. (Baron, Pettit and Slote 1997, 273-280) His focus is upon whether political proposals are virtue-ethically defective, as issuing from or being supported or sustained on account of some vice, or, alternatively, whether they issue from or are supported or sustained by some virtue. As I see it, this overlooks or sidesteps the questions raised by the necessity that a political order in some way take account of the less than fully virtuous among us.

political issues? What *is* it to act and respond well with respect to the issues faced by a political decision-maker?[51] This is more a problem in politics than elsewhere, because in other areas we draw upon a larger fund of experience. A deep feature of a virtue-based approach to ethical issues is that one does not expect to find rules, codifiable in advance, to dictate all of one's steps. Perception of the particular situation and responding appropriately plays an important and ineliminable role.[52] Much of our understanding of the virtues is acquired, shaped and refined in the setting of individual lives and small-group interactions. But when we try to transfer the concepts and practices that have served us well in individual and small-group contexts to apply to issues that impinge upon large groups of mostly anonymous others – that is, to contexts for which our experience, for the most part, has not prepared us – it is not clear that, or how far, or with what qualifications, our concepts apply.

The executive virtues[53] would, no doubt, have a place in any well-lived life in almost any imaginable setting. That is because they are contributory to success in whatever one is doing. But it is hardly enough, in the political realm, to urge courage, ambition, perseverance and the like without saying anything about the causes or goals in the service of which courage, ambition and perseverance are commended.

[51] I am referring primarily to those occupying some public office or official position. There are related questions about the right way to behave as a citizen.

[52] "[T]he whole account of matters of conduct must be given in outline and not precisely, as we said at the very beginning that the accounts we demand must be in accordance with the subject-matter; matters concerned with conduct and questions of what is good for us have no fixity, any more than matters of health. The general account being of this nature, the account of particular cases is yet more lacking in exactness; for they do not fall under any art or set of precepts, but the agents themselves must in each case consider what is appropriate to the occasion, as happens also in the art of medicine or of navigation." (*NE* 1104a 1-9) See also *NE* 1094a 13-27 and McDowell 1996.

[53] I borrow the term from O'Neill 1996, 187-188. The executive virtues "are manifested in deciding on, controlling and guiding action, policies and practices of all sorts." (187)

What is much less clear is what happens to the other excellences of character as they are realized and exhibited appropriately in the political realm. There are two points here. One is that experience with the political realm may require that recognized virtues be qualified in ways that are either not appropriate or not necessary in individual or small-group contexts. This is a special case of the general point that what a virtue amounts to, what it *is* to practice a virtue in a given situation, depends upon what is of importance in that situation, including any other virtues that are called for.[54] The second is that there may be distinctive excellences in the political realm that can only be properly recognized and appreciated through experience. To the extent that either or both of these conditions hold, our understanding of the relevant virtues of political life may be defective and our application of moral concepts to persons and issues involved in the political realm in one way or another inappropriate.

A further important reason turns upon the fact that what is at stake in the political realm is of concern to all citizens. This has a bearing in several ways. First, there is the question of what is to be required, given that not all are virtuous and given that those charged with imposing, administering and enforcing requirements cannot themselves all be counted upon to be virtuous.[55] It may well be that the standards for what it is appropriate to require are different from the standards governing what ought to be done. In fact, it is straightforward that there is a relevant difference here. Insofar as requirements in a political order are associated with sanctions, a necessary condition for

[54] See Chapter Four, note 72.

[55] This is leaving aside the earlier mentioned fact that principles of right conduct are not fully codifiable and therefore not fully codifiable in law. Even for the guidance of fully virtuous people, it would not be possible to make the requirements of the law coincide perfectly with what morally ought to be done.

the proper imposition of a requirement is not just that what is required is something that ought to be done, but also that non-performance is something that ought to be met, or at least is permissibly met, with the imposition of the relevant sanction. If there are any cases in which what is wrong cannot properly be sanctioned, then, to that extent, proper law – what can properly be required – will not be simply a reflection of what morally ought to be done.

5.6 Summary

Many questions remain for exploration. Some focus upon the reach of an instrumental case for eudaemonism and the implications of that for practical reason in general. Others are related to applications of the kind of eudaemonism or virtue ethics sketched here to the political realm. Undoubtedly, there are many more. There is, in any event, no shortage of related issues in need of further investigation and research.

For the present, however, summary of the main conclusions of this chapter may be helpful. It is natural to suppose or to argue that instrumental reasoning cannot bear upon final or ultimate ends, and, if there is no non-instrumental form of practical reason, that final or ultimate ends must have some non-rational source. The supposition or argument, however, is mistaken because the acquisition of ends, even if they are final or ultimate, can serve maieutic objectives which have the role of giving birth to other ends.

The claim that final or ultimate ends *may* be selected or adopted as the upshot of instrumental reasoning stands in need of elaboration. I argue that a pervasive feature of human psychology, the conflict of ends, is a problem, solving which is a maieutic objective, and to which the best solution is the construction of an ultimate end (or embarking upon a corresponding developmental process) in terms of which end-pursuit can be harmonized. The ideal end-point of the developmental process can be identified as comprehensively successful living, or eudaemonia.

Though individuals can be expected to differ substantially in the component ends that will be included for them in comprehensively successful living, we can expect common features as well. These have their basis in a common human nature, which, in our shared world, sets us certain common problems as well as establishing some motivational constraints, in recurrent situation-types with which we are faced, and in the fact that acquiring certain dispositions of intelligent responsiveness to those recurrent situation-types can be expected to serve us well. This is how the excellences of character enter into and qualify comprehensively successful living.

REFERENCES

Ackrill, J. L. 1980. Aristotle on Eudaimonia. In *Essays on Aristotle's Ethics*, edited by A. O. Rorty. Berkeley: University of California Press.

Allais, Maurice. 1990/1979. Criticism of the postulates and axioms of the American School. In *Rationality in Action: Contemporary Approaches*, edited by P. K. Moser. Cambridge: Cambridge University Press.

Allen, Colin, Marc Bekoff, and George Lauder, eds. 1998. *Nature's Purposes: Analyses of Function and Design in Biology*. Cambridge: MIT Press (A Bradford Book).

Annas, Julia. 1993. *The Morality of Happiness*. New York: Oxford University Press.

Aristotle. 1984. *The Complete Works of Aristotle: The Revised Oxford Translation, Bollingen series LXXI:2*. Princeton: Princeton University Press.

_____. 1997. *Politics: Books VII and VIII.* Translated by Richard Kraut (with Commentary). Edited by J. L. Ackrill and L. Judson, *Clarendon Aristotle Series*. Oxford: Clarendon Press.

Arnhart, Larry. 1998. *Darwinian Natural Right: The Biological Ethics of Human Nature*. Edited by D. E. Shaner, *SUNY Series in Philosophy and Biology*. State University of New York Press.

Audi, Robert. 1997. Moral Judgement and Reasons for Action. In *Ethics and Practical Reason*, edited by G. Cullity and B. Gaut. Oxford: Clarendon Press.

Austin, J.L. 1970. Agathon and Eudaimonia in the *Ethics* of Aristotle. In *Philosophical Papers*, edited by J. O. Urmson and G. J. Warnock. London: Oxford University Press.

Baron, Marcia W., Philip Pettit, and Michael Slote. 1997. *Three Methods of Ethics: A Debate*. Malden: Blackwell Publishers.

Bass, Robert. 1994. Choice under Uncertainty [Manuscript].

Bedau, Mark. 1998. Where's the Good in Teleology? In *Nature's Purposes: Analyses of Function and Design in Biology*, edited by C. Allen, M. Bekoff and G. Lauder. Cambridge: MIT Press (A Bradford Book).

Bentham, Jeremy. 1973. An introduction to the principles of morals and legislation. In *The Utilitarians*. Garden City: Anchor Books.

Bratman, Michael E. 1999/1987. *Intention, Plans, and Practical Reason*. In *David Hume Series of Philosophy and Cognitive Science*. Stanford: Center for the Study of Language and Information.

Broadie, Sarah. 1991. *Ethics with Aristotle*. New York: Oxford University Press.

_____. 1987. The problem of practical intellect in Aristotle's ethics. In *Proceedings of the Boston Area Colloquium in Ancient Philosophy*, vol. 3, edited by J. Cleary. Lanham: University Press of America.

Broome, John. 1999a. *Ethics Out of Economics*. Cambridge: Cambridge University Press.

_____. 1999b. 'Utility'. In *Ethics Out of Economics*. Cambridge: Cambridge University Press.

Camerer, Colin. 1995. Individual Rationality. In *The Handbook of Experimental Economics*, edited by J. Kagel and A. Roth. Princeton: Princeton University Press.

Cooper, John M. 1996. Eudaemonism, the Appeal to Nature, and "Moral Duty" in Stoicism. In *Aristotle, Kant and the Stoics: Rethinking Happiness and Duty*, edited by S. Engstrom and J. Whiting. Cambridge: Cambridge University Press.

Cosmides, Leda, and John Tooby. 1992. Cognitive Adaptations for Social Exchange. In *The Adapted Mind*, edited by J. H. Barkow, L. Cosmides and J. Tooby. New York: Oxford University Press.

Cummins, Robert. 1998. Functional Analysis. In *Nature's Purposes: Analyses of Function and Design in Biology*, edited by C. Allen, M. Bekoff and G. Lauder. Cambridge: MIT Press (A Bradford Book).

Darwall, Stephen. 1983. *Impartial Reason*. Ithaca: Cornell University Press.

_____. 1997. Reasons, Motives, and the Demands of Morality: An Introduction. In *Moral Discourse and Practice*, edited by S. Darwall, A. Gibbard, and P. Railton. New York: Oxford University Press.

Dawes, Robyn M. 1988. *Rational Choice in an Uncertain World.* Fort Worth: Harcourt Brace College Publishers.

Dawkins, Richard. 1976. *The Selfish Gene*. Oxford: Oxford University Press.

_____. 1982. *The Extended Phenotype: The Gene as the Unit of Selection*. Oxford: Oxford University Press.

Dennett, Daniel C. 1995. *Darwin's Dangerous Idea*. New York: Simon and Schuster.

Dworkin, Ronald. 1978. *Taking Rights Seriously*. Cambridge: Harvard University Press.

Ellsberg, Daniel. 1990/1961. Risk, ambiguity, and the Savage axioms. In *Rationality in Action: Contemporary Approaches*, edited by P. K. Moser. Cambridge: Cambridge University Press.

Elster, Jon. 1979. *Ulysses and the Sirens: Studies in Rationality and Irrationality*. Cambridge: Cambridge University Press.

Finnis, John. 1980. *Natural Law and Natural Rights*. Oxford: Clarendon Press.

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

_____. 2002. *The Problem of the Soul*. New York: Basic Books.

Foot, Philippa. 1978. Virtues and Vices. In *Virtues and Vices and Other Essays in Moral Philosophy*. Berkeley: University of California Press.

Found, Peter. 2001. *Never Judge a Dutch Book by its Cover*. Ph.D Dissertation. Bowling Green: Bowling Green State University.

Frankfurt, Harry. 1969. Alternate Possibilities and Moral Responsibility. *Journal of Philosophy* 65: 829-833.

Friedman, David. 1996. *Hidden Order: The Economics of Everyday Life*. New York: Harper Business.

Gauthier, David P. 1986. *Morals by Agreement*. Oxford [Oxfordshire]; New York: Clarendon Press; Oxford University Press.

Gould, Steven Jay. 1991. Exaptation: A Crucial Tool for an Evolutionary Psychology. *Journal of Social Issues* 47 (3): 43-65.

Hampton, Jean and Richard Healey. 1998. *The Authority of Reason*. Cambridge: Cambridge University Press.

Hardie, W. F. R. 1968. The Final Good in Aristotle's Ethics. In *Aristotle: A Collection of Critical Essays*, edited by J. M. E. Moravcsik. Notre Dame: University of Notre Dame Press.

Hare, R. M. 1981. *Moral Thinking*. Oxford: Oxford University Press.

Harman, Gilbert. 1977. *The Nature of Morality: An Introduction to Ethics*. New York: Oxford University Press.

_____. 2000. The Nonexistence of Character Traits. *Proceedings of the Aristotelian Society 1999-2000*, 100: 223-226.

_____. 2003. No Character or Personality. *Business Ethics Quarterly* 13: 87-94.

Harman, Gilbert, and Judith Jarvis Thomson. 1996. *Moral Relativism and Moral Objectivity, Great Debates in Philosophy*. Cambridge, Mass., USA: Blackwell.

Hart, H. L. A. 1984. Are There Any Natural Rights? In *Theories of Rights*, edited by J. Waldron. Oxford: Oxford University Press.

Heap, Shaun Hargreaves, Martin Hollis, Bruce Lyons, Robert Sugden, and Albert Weale. *The Theory of Choice: A Critical Guide*. Cambridge: Blackwell.

Hurka, Thomas. 1993. *Perfectionism, Oxford ethics series*. New York: Oxford University Press.

Hurley, Susan L. 1989. *Natural Reasons: Personality and Polity*. New York: Oxford University Press.

Irwin, T. H. 1980. The Metaphysical and Psychological Basis of Aristotle's Ethics. In *Essays on Aristotle's Ethics*, edited by A. O. Rorty. Berkeley: University of California Press.

Kahneman, Daniel and Amos Tversky. 1990. Prospect theory: an analysis of decision under risk. In *Rationality in Action: Contemporary Approaches*, edited by P. K. Moser. Cambridge: Cambridge University Press.

Kant, Immanuel. 1960. *Religion Within the Limits of Reason Alone*. Translated by Theodore M. Greene, Hoyt H. Hudson. New York: Harper and Row.

_____. 1997. *Critique of Practical Reason*. Translated by Mary Gregor. Edited by K. Ameriks and D. M. Clark, *Cambridge Texts in the History of Philosophy*.

Cambridge: Cambridge University Press.

_____. 1998. *Groundwork of the Metaphysics of Morals*. Translated by Mary J. Gregor (with Introduction by Christine M. Korsgaard), *Cambridge texts in the history of philosophy*. Cambridge: Cambridge University Press.

Kavka, Gregory. 1986. *Hobbesian Moral and Political Theory*. Edited by M. Cohen, *Studies in Moral, Political, and Legal Philosophy*. Princeton: Princeton University Press.

Korsgaard, Christine M. 1996a. Skepticism about Practical Reason. In *Creating the Kingdom of Ends*. Cambridge: Cambridge University Press.

_____. 1996b. *The Sources of Normativity*. Cambridge; New York: Cambridge University Press.

_____. 1997. The Normativity of Instrumental Reason. In *Ethics and Practical Reason*, edited by G. Cullity and B. Gaut. Oxford: Clarendon Press.

Lippmann, Walter. 1957. *A Preface to Morals*. New York: Time-Life Books.

Locke, John. 1975. *An Essay Concerning Human Understanding*. Edited by P. H. Nidditch. Oxford: Clarendon Press.

Long, Roderick T. 1993. Living by Degrees: Ethics for Discursive Animals [Manuscript].

Luce, R. Duncan and Howard Raiffa. 1985/1957. *Games and Decisions: Introduction and Critical Survey*. New York: Dover Publications.

Machan, Tibor R. 1975. *Human Rights and Human Liberties: A Radical Reconsideration of the American Political Tradition*. Chicago: Nelson Hall.

MacIntyre, Alasdair C. 1984. *After Virtue: A Study in Moral Theory* [Second edition]. Notre Dame: University of Notre Dame Press.

Mack, Eric. 1971. How to Derive Ethical Egoism. *Personalist* (Autumn): 735-743.

Mackie, J. L. 1977. *Ethics: Inventing Right and Wrong, Pelican books: Philosophy*. Harmondsworth; New York: Penguin.

McClennen, Edward F. 1990. *Rationality and Dynamic Choice: Foundational Explorations*. Cambridge: Cambridge University Press.

McDowell, John. 1996. Deliberation and Moral Development in Aristotle's Ethics. In *Aristotle, Kant, and the Stoics: Rethinking Happiness and Duty*, edited by S. Engstrom and J. Whiting. Cambridge: Cambridge University Press.

Mill, John Stuart. 1965. Utilitarianism. In *Mill's Ethical Writings*, edited by J. B. Schneewind. New York: Collier Books.

Miller, Fred Dycus. 1995. *Nature, Justice, and Rights in Aristotle's Politics*. Oxford; New York: Clarendon Press; Oxford University Press.

Millgram, Elijah. 1997. *Practical Induction*. Cambridge: Harvard University Press.

Millikan, Ruth Garrett. 1984. *Language, Thought, and Other Biological Categories: New Foundations for Realism*. Cambridge: MIT Press.

_____. 1998. In Defense of Proper Functions. In *Nature's Purposes: Analyses of Function and Design in Biology*, edited by C. Allen, M. Bekoff and G. Lauder. Cambridge: MIT Press (A Bradford Book).

Mises, Ludwig von. 1963. *Human Action: A Treatise on Economics* [Third revised edition]. Chicago: Henry Regnery Company.

Murray, Charles. 1988. *In Pursuit of Happiness and Good Government.* New York: Simon and Schuster.

Nagel, Ernest. 1968. The structure of teleological explanations. In *The Philosophy of Science*, edited by P. H. Nidditch. Oxford: Oxford University Press.

Nagel, Thomas. 1980. Aristotle on *Eudaemonia.* In *Essays on Aristotle's Ethics*, edited by A. O. Rorty. Berkeley: University of California Press.

Neander, Karen. 1998. Functions as Selected Effects: The Conceptual Analyst's Defense. In *Nature's Purposes: Analyses of Function and Design in Biology*, edited by C. Allen, M. Bekoff and G. Lauder. Cambridge: MIT Press (A Bradford Book).

Nozick, Robert. 1981. *Philosophical Explanations.* Cambridge: Belknap Press of the Harvard University Press.

_____. 1993. *The Nature of Rationality.* Princeton: Princeton University Press.

_____. 1997. On Austrian Methodology. In *Socratic Puzzles.* Cambridge: Harvard University Press.

Nussbaum, Martha Craven. 1986. *The Fragility of Goodness: Luck and Ethics in Greek Tragedy and Philosophy.* Cambridge; New York: Cambridge University Press.

_____. 1993. Non-relative virtues: an Aristotelian approach. In *The Quality of Life,* edited by M. C. Nussbaum and A. Sen. Clarendon: Oxford University Press.

_____. 1994. *The Therapy of Desire: Theory and Practice in Hellenistic Ethics.* Princeton: Princeton University Press.

_____. 2001. *Upheavals of Thought: The Intelligence of Emotions.* Cambridge: Cambridge University Press.

O'Neill, Onora. 1996. *Towards Justice and Virtue: A Constructive Account of Practical Reasoning.* Cambridge: Cambridge University Press.

Peirce, Charles S. 1957/1878. The Doctrine of Chances. In *Essays in the Philosophy of Science*, edited by Vincent Tomas. Indianapolis: Bobbs-Merrill.

Pinker, Steven. 1997. *How the Mind Works.* New York: W. W. Norton and Co.

_____. 2002. *The Blank Slate: The Modern Denial of Human Nature.* New York: Viking.

Piper, Adrian M. S. 1986. Instrumentalism, Objectivity, and Moral Justification. *American Philosophical Quarterly* 23 (Number 4, October): 373-381.

Popper, Karl R. 1965. *Conjectures and Refutations: The Growth of Scientific Knowledge.* New York: Harper and Row.

Rasmussen, Douglas B., and Douglas J. Den Uyl. 1991. *Liberty and Nature: An Aristotelian Defense of Liberal Order*. La Salle: Open Court.

Raz, Joseph. 1986. *The Morality of Freedom*. Oxford: Clarendon Press.

Rawls, John. 1971. *A Theory of Justice*. Cambridge: Belknap Press.

Richards, R. J. 1995. A Defense of Evolutionary Ethics. In *Issues in Evolutionary Ethics*, edited by P. Thompson. Albany: State University of New York Press.

Root, Michael. 1993. *Philosophy of Social Science*. Oxford: Blackwell Press.

Russell, Bertrand. 1955. *Human Society in Ethics and Politics*. New York: Simon and Schuster.

Sartre, Jean-Paul. 1975/1946. Existentialism is a Humanism. In *Existentialism from Dostoevsky to Sartre*, edited by W. Kaufmann. New York: New American Library.

Savage, Leonard J. 1972. *The Foundations of Statistics* [Second revised edition]. New York: Dover.

Scanlon, T. M. 1998. *What We Owe to Each Other*. Cambridge: Belknap Press of Harvard University Press.

Schmidtz, David. 1995. *Rational Choice and Moral Agency*. Princeton, N.J.: Princeton University Press.

Simon, Herbert A. 1990/1983. Alternative visions of rationality. In *Rationality in*

Action: Contemporary Approaches, edited by P. K. Moser. Cambridge: Cambridge University Press.

_____. 1996/1969 [Third edition]. *The Sciences of the Artificial.* Cambridge: The MIT Press.

Sober, Elliott. 1993. *Philosophy of Biology*. Edited by N. Daniels and K. Lehrer, *Dimensions of Philosophy*. Boulder: Westview Press.

Solomon, Robert C. 1976. *The Passions: The Myth and Nature of Human Emotion.* Garden City: Anchor Books.

Sorabji, Richard. 1980. *Necessity, Cause and Blame: Perspectives on Aristotle's Theory*. Ithaca: Cornell University Press.

Sterelny, Kim. 1995. The Adapted Mind. *Biology and Philosophy* 10: 365-380.

Stevenson, Charles L. 1944. *Ethics and Language.* New Haven: Yale University Press.

Strauss, Leo. 1953. *Natural Right and History*. Chicago: University of Chicago Press.

Verbeek, Bruno. 1999. Rationality, Consequentialism, and the Problem of Relevant Description of Outcomes. [Manuscript].

Wallace, James. 1978. *Virtues and Vices*. Ithaca: Cornell University Press.

Wild, John Daniel. 1953. *Plato's Modern Enemies and the Theory of Natural Law*. Chicago: University of Chicago Press.

Williams, Bernard. 1990. Internal and external reasons. In *Rationality in Action:*

Contemporary Approaches, edited by P. K. Moser. Cambridge: Cambridge University Press.

Wilson, James Q. 1993. *The Moral Sense*. New York: The Free Press.

Wright, Larry. 1998. Functions. In *Nature's Purposes: Analyses of Function and Design in Biology*, edited by C. Allen, M. Bekoff and G. Lauder. Cambridge: MIT Press (A Bradford Book).

Wright, Robert. 1994. *The Moral Animal*. New York: Vintage Books.